PREDICTIVE MODELING OF AVIAN INFLUENZA IN WILD BIRDS

A

DISSERTATION

Presented to the Faculty

of the University of Alaska Fairbanks

in Partial Fulfillment of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

By

Keiko A. Herrick, B.A., M.A.

Fairbanks, Alaska

May 2013

ABSTRACT

Over the past 20 years, highly pathogenic avian influenza (HPAI), specifically Eurasian H5N1 subtypes, caused economic losses to the poultry industry and sparked fears of a human influenza pandemic. Avian influenza virus (AIV) is widespread in wild bird populations in the low-pathogenicity form (LPAI), and wild birds are thought to be the reservoir for AIV. To date, however, nearly all predictive models of AIV focus on domestic poultry and HPAI H5N1 at a small country or regional scale. Clearly, there is a need and an opportunity to explore AIV in wild birds using data-mining and machine-learning techniques.

I developed predictive models using the Random Forests algorithm to describe the ecological niche of avian influenza in wild birds. In "*Chapter 2 - Predictive risk modeling of avian influenza around the Pacific Rim*", I demonstrated that it was possible to separate an AIV-positivity signal from general surveillance effort. Cold winters, high temperature seasonality, and a long distance from coast were important predictors. In "*Chapter 3 - A global model of avian influenza prediction in wild birds: the importance of northern regions*", northern regions remained areas of high predicted occurrence even when using a global dataset of AIV. In surveillance data, the percentage of AIV-positive samples is typically very low, which can hamper machine-learning. For "*Chapter 4 - Modeling avian influenza with Random Forests: under-sampling and model selection for unbalanced prevalence in surveillance data*" I wrote custom code in R statistical programming language to evaluate a balancing algorithm, a model selection algorithm, and an under-sampling method for their effects on model accuracy. Repeated random

sub-sampling was found to be the most reliable way to improved unbalanced datasets. In these models cold regions consistently bore the highest relative predicted occurrence scores for AIV-positivity and describe a niche for LPAI that is distinct from the niche for HPAI in domestic poultry. These studies represent a novel, initial attempt at constructing models for LPAI in wild birds and demonstrated high predictive power.

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

Page

LIST OF ADDITIONAL MATERIALS

LIST OF APPENDICES

DEDICATION

This work is dedicated to as-of-yet-unnamed Baby Akasofu-Herrick: the impetus and motivation to finish and move on to the next phase of Life. See you in August!

ACKNOWLEDGEMENTS

I am truly grateful and indebted to many people for their support and guidance through the past several years. In particular, I thank DR. FALK HUETTMANN for his unwavering support. I am inspired by his passion for his work and his boundless enthusiasm. Quality wins! I thank DR. BARBARA TAYLOR for sticking with me through the years, for her invaluable guidance, and Canadian humor. I thank my committee members: DR. STEFFI ICKERT-BOND and DR. ERIC BORTZ for their patience, positivity, and support. I thank all my previous advisors: DR. JON RUNSTADLER, DR. TOM MARR, and DR. ABEL BULT-ITO. Thank you, Abel, for setting me on this road and for your friendship through the years. I thank Biology Department Chair DR. CHRISTA MULDER and Dean of CNSM DR. PAUL LAYER for comments and critiques on this dissertation, which helped to make it stronger. Thanks to DR. ABBY POWELL and the Scientific Writing class for help with Chapter 3. I have been honored to work with a multitude of colleagues in the Bult-Ito, Marr, Runstadler, and EWHALE labs. Thank you all for everything; I am fortunate to have such friends! I would like to express my great appreciation for CAROL PISER, DEANNA FITZGERALD, CATHY GRISETO, and CAROL COLP, who were advisors, advocates, protectors, and friends. I am eternally grateful for the support and love of my family: ROBERT HERRICK, SYUN-ICHI & EMIKO AKASOFU, KEN-ICHI & MASAE AKASOFU, and of course, BUZZ and ASTRO. I have too many friends to list, but in particular, thanks to VIVICA GENAUX, BRIAN & SARA YOUNG, and DANIELLE & MATT DILLON.

CHAPTER 1

General Introduction

Avian influenza virus, transmission, and pandemic potential

Influenza A virus is an orthomyxovirus with a genome of negative-stranded RNA made up of 8 segments. Avian influenza virus (AIV) is a member of this *Orthomyxoviridae* family and is classified by subtype, which is designated by hemagglutinin (HA) and neuraminidase (NA) surface protein antigen names (such as H5N1). To date, 17 HA subtypes and 10 NA subtypes have been isolated and identified and nearly every HA/NA combination has been found in wild birds [1-3]. This dissertation may refer to strains within each subtype, which can mean specific viral isolates, specific genotypes, or broader classes of genotypes within a subtype. AIV is also classified as highly pathogenic avian influenza (HPAI) or low pathogenicity avian influenza (LPAI). HPAI and LPAI strains tend to differ between their HA0 cleavage site amino acid sequence with certain characteristics that permit high cleavability that facilitates infection [4]. According to the World Organization for Animal Health (OIE), any virus that is found to have a sequence similar to HPAI strains is considered to be a notifiable disease [5] that is potentially highly pathogenic. However, HPAI or LPAI classification is ultimately based on a strain's lethality to chickens. Strains are designated HPAI if they cause greater than 75% mortality (where $n \geq 8$) within 10 days of inoculation [5], although the mortality rate is frequently 100% within a day or two [6]. While LPAI is nearly ubiquitous in wild birds and thought to be distributed worldwide [7], only certain H5 and H7 subtypes have been

identified as HPAI (including H5N1, H7N7, and H7N3). Most strains of these subtypes are LPAI and few H5 and H7 strains are highly pathogenic.

The HPAI H5N1 outbreaks in the 2000's raised concerns over the potential for deadly "bird flu" outbreaks in humans. Indeed, wild birds contributed gene segments to influenza viruses that caused the four deadliest human influenza pandemics (1918, 1957, 1968, 2009) [8]. While human to human transmission of HPAI H5N1 is extremely rare [9], humans have contracted it from exposure to sick birds [10-14]. Cases where highly pathogenic H5N1 was directly transmitted from chicken to human were first recorded in 1997 when H5N1 was isolated from a 3 year old boy in Hong Kong, who subsequently died of influenza-related complications [12]. Clinical disease has not been reported among poultry workers, but researchers have documented seroconversion, or the development of HPAI H5N1-specific antibodies [15, 16], which is a sign of exposure to the virus. More recently, H5N1 was isolated from humans in Vietnam in 2003 [13], a cluster of Indonesian H5N1 cases was reported in 2006 [11], and as of this writing, the World Health Organization (WHO) lists over 600 human cases with a mortality rate of over 50% [14]. H5N1 is not the only potential threat to human health. Other subtypes have raised concerns over new pandemics [17, 18] with reports of human infection by subtypes such as H9N2 [10, 19] and H11N9 [20].

A pandemic begins when three conditions are met: 1) a viral subtype novel to humans emerges, 2) this subtype infects humans and is highly pathogenic, and 3) the virus in its novel form is transmitted between humans [21]. Humans have no immunity to this novel strain and if it is transmitted as easily as seasonal influenza, the outcome is

potentially devastating. While human to human transmission of H5N1 has not been verified, the 2009 form of H1N1 ("swine flu") has already made this step. The H1N1 pandemic, which began in 2009 [22], presents a novel combination of viral proteins from Eurasian swine virus, classical swine virus, and a triple reassortment (bird, human, swine) of North American origin. While swine serve as the intermediary to human infection, all these strains were originally seeded by wild birds (reviewed in [8]).

AIV has demonstrated the potential to infect many types of vertebrates. In the wild and in the laboratory, AIV has been isolated most commonly from dabbling ducks (subfamily *Anatinae*), shorebirds (Charadriiformes excluding web-footed waders), and domestic Galliformes (chickens, turkeys, quail, etc.), but also rails (Gruiformes) [2], passerines [23], parrots (Psittaciformes) [24], American kestrel (*Falco sparverius*) [25], ostriches [26], owls, doves, and storks [27], to name a few. Mammals that have been infected with AIV in both experimental and non-experimental settings include domestic cats [28], dogs [29], pigs [30], mice and ferrets [31], tigers [32], stone marten (*Martes foina*) [33], and marine mammals [34]. Some of these species live in close quarters with humans as pets, commensal pests, or in high-density livestock settings with chickens, such as farms [35] or live bird markets [36], increasing the potential of AIV transmission to humans. If waterfowl transmit AIV to chickens, then mammals such as pigs, barn cats, or rodents could contract the virus and in turn, infect humans.

Chickens and other Galliformes such as turkey (*Meleagris gallopova*) and quail (*Coturnix coturnix japonica*) are particularly susceptible to infection by AIV and serve as novel or aberrant host species in which LPAI strains rapidly evolve into more pathogenic

ones [37, 38]. Chickens are a species of agricultural importance often kept in high densities or in close quarters with humans, increasing the potential for transmission of AIV to humans. Different haplotypes of chicken demonstrate resistance and susceptibility to H5N1 [39, 40]. Rural chickens display high haplotype diversity compared to commercial lines, perhaps from interbreeding with wild-type fowl [41]. This diversity appears to confer high levels of immunity to AIV for some haplotypes, including immunity to HPAI H5N1 [39, 40]. Commercial European breeds display intermediate levels of genetic diversity [42]. Meat chickens, selected for growth capacity, may suffer immune-compromise as a trade-off for rapid growth [43]. When wild birds intermingle with and infect domestic poultry, egg and meat production can be impaired and may necessitate culling of entire flocks to control the spread of disease [44, 45]. Most domestic outbreaks come from new introductions by wild birds, not from virus endemic to poultry [46].

The economic impact of AIV to poultry farming in the United States has been as high as $149M for a single incident [47]. H5N1, which is so far restricted to the Eastern Hemisphere, is responsible for the death (by disease or culling) of over 400 million domestic poultry at a cost of over $20 billion USD. [48]. The U.S. Department of Agriculture (USDA), Animal and Plant Health Inspection Service (APHIS), the Department of the Interior (DOI), and the Department of Health and Human Services (HHS) maintain a National Highly Pathogenic Avian Influenza (HPAI) Response Plan [49], which was created to complement industry and state/regional plans for humans and livestock . This plan does not appear to address free-range or pasture-raised poultry,

which are of concern because they have access to the outdoors, where they have a greater potential to come in contact or share water with wild waterfowl than do conventional, indoor poultry. APHIS maintains a website ("Biosecurity for Birds"; http://www.aphis.usda.gov/animal_health/birdbiosecurity/) targeted toward backyard poultry hobbyists that introduces biosecurity concepts and describes clinical signs of avian influenza [50].

Once AIV is transmitted to chickens from wild birds, it evolves rapidly [51]; low pathogenicity strains, which cause no clinical symptoms and replicate poorly in chickens, can grow increasingly virulent and convert to a highly pathogenic form simply by being passed from bird to bird [38, 52-56]. While many of these studies were done with H5 and H7 strains of AIV, H11 [53] and H9 [57] strains also demonstrate rapid adaptation and increased pathogenicity with passage through chickens. Similarly, vaccination appears to create selective pressure on the virus such that immunized birds shed highly pathogenic antigenic variants (with surface protein mutations that are not recognized by antibodies) against which they have low immunity [58]. This phenomenon is not restricted to chickens and AIV: it has been reported in experimentally AIV-infected turkeys (*Meleagris gallopavo*) [56]; domestic ostrich (*Struthio camelus*), in which an LPAI strain from wild birds mutated into an HPAI strain [26]; and an avirulent strain of Newcastle disease virus, which became highly pathogenic via experimental passage through chickens [59].

Waterfowl are typically considered resistant to AIV and in the laboratory appear to tolerate infection by LPAI viruses without obvious clinical signs of infection [60, 61].

Even when infected with highly pathogenic strains of H5 or H7, ducks displayed a slight increase in body temperature [62] or mild upper respiratory symptoms despite viral infection in organs such as spleen, liver, kidney, and brain [63]. Those that survived HPAI infection continued to shed large quantities of virus, which was less pathogenic to naïve ducks, but still highly pathogenic and lethal to chickens [45] and presumably to other species. Experimentally infected mallards (*Anas platyrhynchos*) at normal body weight were more susceptible to infection and shed more virus than ducks manipulated to reduced body weight [64]. These counterintuitive results suggest the virus has evolved a survival strategy in ducks to optimize its spread and transmission. Even the H5N1 strains that emerged in 2002 appear to be decreasing in pathogenicity to ducks. While these strains initially killed large numbers of wild waterfowl in Asia and were highly pathogenic to ducks in the laboratory [65], these strains have evolved to become non-pathogenic to ducks, while remaining highly pathogenic to poultry [45]. AIV may have a more noticeable effect on wild birds more than on laboratory birds, Bewick swans (*Cygnus columbianus bewickii*) infected with LPAI displayed impaired foraging and migratory performance [66] and LPAI infection increased staging time and decreased body mass in wild mallards (*Anas platyrhynchos*) [67]. It appears that waterfowl are not resistant, but rather readily infected with AIV. Furthermore, ducks in a laboratory setting with ample food and comfortable living conditions may display greater tolerance to AIV infection and fewer clinical symptoms than wild waterfowl undergoing the stresses of migration or seasonal breeding.

Modeling AIV

Remote sensing data and Geographic Information Systems (GIS) technology have proven useful in the management of infectious disease [68-70]. When combined with advanced statistical prediction algorithms, GIS has the power to integrate diverse types of data to describe and detect complex patterns, then quantitatively model and extrapolate the findings. Additional surveillance can then assess the "ground truth" of model predictions, thus improving their predictive accuracy. Ecological niche modeling is used more commonly in parasite-host cases such as Chagas disease [68], Lyme disease [70], malaria [71], and West Nile virus [72]. The study of Lyme disease, for example, has made extensive use of GIS technologies in predicting risk based on habitat suitability of Lyme ticks [73], identifying environmental factors correlated with Lyme disease [74], and forecasting the expansion of tick-suitable habitat based on projected climate change [75].

AIV modeling can be generally divided into a) models of the mechanics of transmission, b) regional-scale studies, and c) trans-regional studies. Research into transmission mechanics includes a number of studies (reviewed in [76]) that have used remote sensing data to identify individual environmental indicators correlated with the survival of AIV in water, such as temperature and salinity. While this review acknowledged the role of water in AIV transmission is still poorly understood, it underscored the availability of technologies such as GIS in analyzing and identifying environmental factors that may play a role in the spread of AIV. Eskici and Turkgulu ([77]) ran several mathematical (non-geographical) scenarios of HPAI pandemic in which chickens became infected by wild ducks and in turn infected humans. These scenarios

accounted for viral mutation and tested the effects of management actions such as culling and quarantine. Poultry density was found to be the most important factor in the emergence of an epidemic. Interestingly, the transmission of HPAI to poultry by wild birds was not found to change the dynamics of spread other than to start the process earlier by skipping the LPAI to HPAI mutation step. Kilpatrick et al. ([78]) constructed a global prediction model of H5N1 spread based on: 1) the number of birds entering a country via migration or bird trade (poultry, pet, and wild); 2) the likelihood they would have picked up AIV prior to entering the focal country; and 3) the number of days these birds were predicted to shed virus [78]. This method predicts that HPAI H5N1 is more likely to enter the Western Hemisphere by the poultry trade rather than via migratory birds. The number of birds entering the Americas through the poultry, pet, and wild bird trade from H5N1-affected countries is more likely than that of birds migrating between affected countries and the Americas.

Regional models focus almost exclusively on HPAI H5N1. Ecological niche models using the Genetic Algorithm for Rule Set Production (GARP) were able to identify risk factors for HPAI H5N1 on a country-wide scale [79, 80]. In Nigeria and West Africa, important predictors included savannah and woodland habitats and a relatively dry climate with high temperature seasonality [80]. The likelihood of outbreak was not correlated with the density of backyard chickens, and I speculate that this may be due to the high haplotype diversity that is often present in mixed rural breeds [39, 40]. In India, important predictors for HPAI H5N1 included low slope angle (which allows for standing bodies of water), high variation in greenness, a mean annual temperature range

of 21-26 °C, and a high human population density (>100 persons/km^2) [79]. Gilbert et al. have used multiple logistic and linear regression models to identify important predictor variables for HPAI H5N1: in Thailand, the predictors most strongly associated with HPAI H5N1 in domestic poultry were the presence of free-grazing ducks, followed by poultry density, and land elevation [81]; in a larger Southeast Asia model, human population density, rice cropping intensity, and free-grazing ducks were important factors [82]. There is some difficulty in generalizing the results from ecological niche models and regression models as they each use a different, limited set of predictors.

Most other regional scale studies are spatiotemporal analyses using infected farms or communities as the infectious unit rather than individual birds. Pfeiffer et al. ([83]) examined spatiotemporal patterns in three waves of HPAI H5N1 outbreak across Vietnam and identified the mean distance to nearest population, population density, high percentage of land used for aquaculture or rice cultivation, as well as domestic poultry density (both ducks and chickens) as contributing factors. Cecchi et al. ([84]) determined that outbreaks in Nigeria in 2006 probably originated from domestic poultry in rural areas near wetlands frequented by migratory waterfowl. A spatial analysis of infected farms in Thailand determined that environmental conditions or land cover characteristics did not contribute to H5N1 outbreaks. The majority of outbreaks were caused by farm-to-farm transmission stemming from a few original cases [85]. This last study underscores the limited scope of HPAI H5N1 studies. While such studies are important in the management of HPAI H5N1, and emphasize good biosafety practice in reducing

transmission to other farms, their findings cannot be generalized to LPAI transmission in wild birds.

Large-scale, cross-regional studies are more appropriate for studying wild birds because migratory waterfowl cover large distances in their yearly migration and are not bounded by artificial country borders. Spatiotemporal analysis of HPAI H5N1 spread from central Asia to Eastern Europe was more consistent with duck migration patterns than anthropogenic factors, such as trade routes [86]. The largest niche model I have found defines an "agro-ecological" niche for HPAI H5N1 worldwide based on 14 climatic, agricultural, and socio-economic factors [87]. Consistent with previous models, this model identified human population, duck, and chicken density as factors contributing to the spread of HPAI H5N1. This model is unique in that it incorporates purchasing power per capita, or relative wealth of a country. While increased wealth usually covaries with increased animal hygiene, it also covaries with increased agricultural intensification, which has been correlated with HPAI H5N1 outbreak. The only model so far that includes LPAI and focuses on birds other than domestic poultry is the spatial regression model by Fuller et al. (2010, [23]), which predicted AIV across the continental United States from wild bird surveillance data. An early thaw date, low minimum temperatures, and high percentage of agricultural land were important predictors. This study is unique in that it found higher prevalence of AIV in passerines than in other bird orders studied.

Specific aims

This thesis details my work in data mining AIV information and constructing predictive maps of avian influenza in wild birds. Chapter 2 "*Predictive risk modeling of avian*

influenza around the Pacific Rim" addresses the question of whether it is possible to define an ecological niche model and construct a predictive map for a system as complicated as avian influenza in wild birds. AIV surveillance projects find that the virus predominantly infects dabbling ducks and shorebirds, two orders that make long distance, yearly migrations between summer breeding grounds and wintering areas. Using the machine-learning ensemble algorithm Random Forests [88], I constructed a model based on georeferenced bird data and laboratory data for AIV-positive and AIV-negative samples, then defined a niche based on anthropogenic and bioclimatic variables. A clear signal, distinct from overall surveillance effort, emerged for AIV-positive samples demonstrating that this predictive model was not simply a species distribution model for ducks and shorebirds. The study area, which extends from longitude 29.5 to -75.17, and from latitude 76.5 to -44.5, is presented in Figure 1.1 along with bird sampling locations. This chapter was originally published in CODATA Germany [89], and has been re-written for clarity. In addition, a table of bird species collected, the number collected, and the AIV prevalence by species for the original database is included.

Chapter 3 "*A global model of avian influenza prediction in wild birds: the importance of northern regions*" investigates whether an ecological niche model for AIV in wild birds can be constructed on a near-global scale. Using publically available, curated wild bird data compiled by an international group of contributors, I constructed an ecological niche model for LPAI in wild birds that included all continents except for Antarctica. One of the benefits of using a data-mining algorithm such as Random Forests is that it is resistant to noise from extraneous, non-contributing predictors. Unlike GARP

and regression, a large number of predictors can easily be included and tested. The finding that northern regions displayed high predicted occurrence underscores the necessity of taking host species' natural range into account, and that modeling in low-latitude areas may not be representative of AIV. Surveillance data have made great contributions to the identification of viral subtypes, infected bird species, and genetic sequence data for AIV. This chapter demonstrates that surveillance data are valuable and useful for machine-learning and predictive modeling purposes as well. The study area and sampling locations for this model are presented in Figure 1.2.

Chapter 4 "*Modeling avian influenza with Random Forests: under-sampling and model selection for unbalanced prevalence in surveillance data*" addresses the highly imbalanced prevalence that is commonly found in wild bird surveillance data. As highly imbalanced data can decrease model accuracy, I evaluate the effectiveness of balancing and model selection techniques on two databases from independent wild bird AIV surveillance projects. The effects of collection in "traditional" surveillance locations year after year were found to have a detrimental effect on predictive modeling. Overly-represented locations tend to skew distributions, resulting in findings than cannot be extrapolated outside the collection location. A down-sampling technique is applied to such a database and the improvement in model accuracy is evaluated. In this chapter, I implemented these algorithms in code written for the R package randomForest. The study area, which extends from longitude 29.5 to -75.17, and from latitude 76.5 to -44.5, is presented in Figure 1.3 along with bird sampling locations from both the Alaska Asia

Avian Influenza Research database (2005-2010) and Canada's Inter-agency Wild Bird Influenza survey database.

Metadata for each dataset as well as metadata for the global predictor variable layers used in these three chapters are included as Appendix A and Appendix B, respectively. The datasets and collected predictor variables themselves are included as a data chapter for this dissertation.

FIGURES

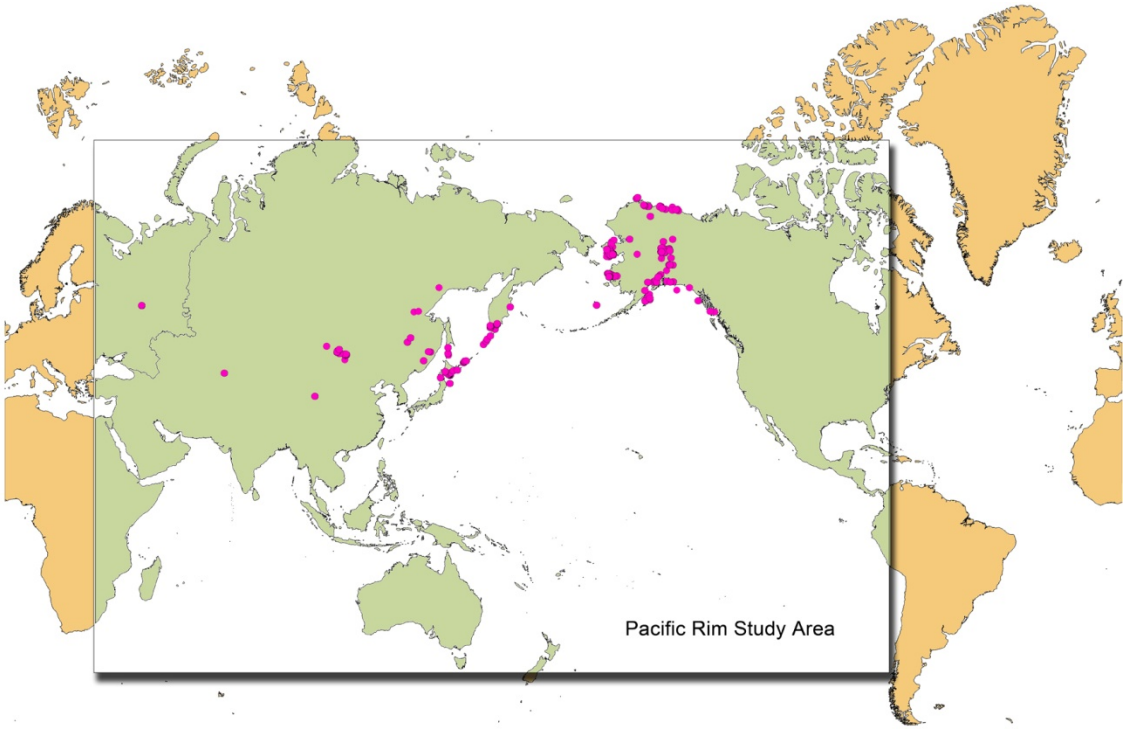

Figure 1.1. Pacific Rim study area and wild bird surveillance locations. These data are from the Alaska Asia Avian Influenza Research group 2005-2007 dataset used in Chapter 2. The bounding rectangle of the study area is represented by the green region within the box. The points indicate where sampling for avian influenza virus (AIV) occurred and include both AIV-positive and AIV-negative points.

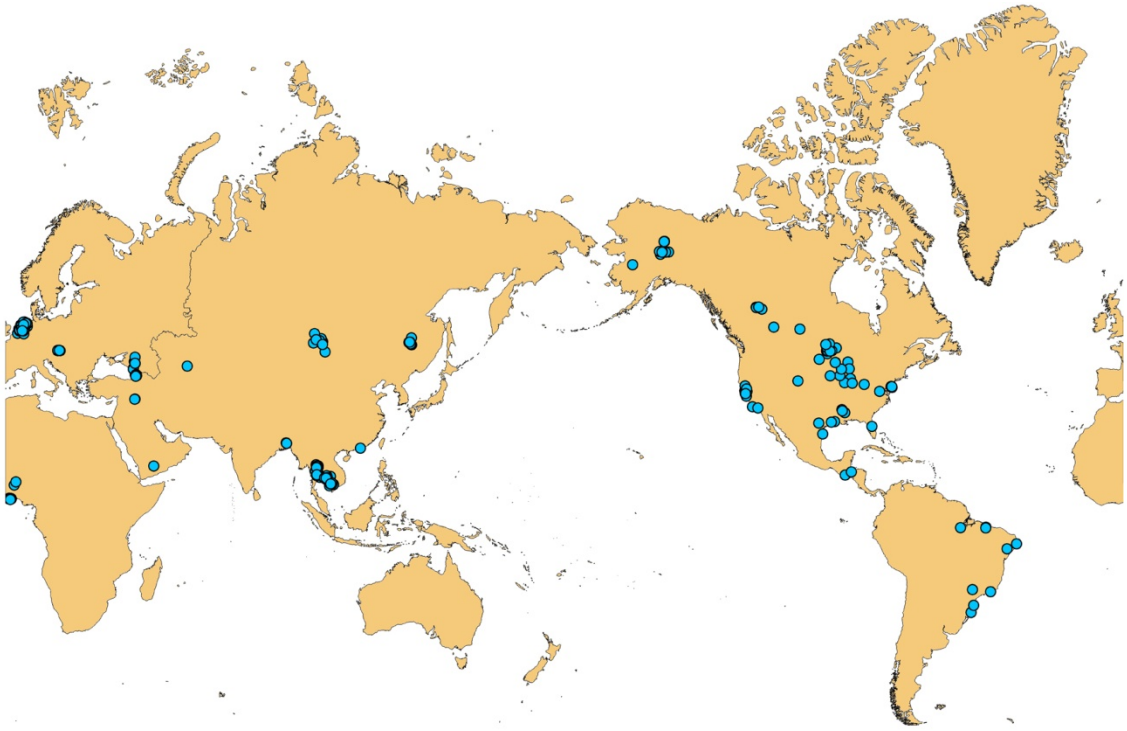

Figure 1.2. Global study area and wild bird surveillance locations. These data are from the NIH Influenza Research Database (fludb.org) used in Chapter 3. The study area excludes Antarctica. The points indicate where sampling for avian influenza virus (AIV) occurred and include both AIV-positive and AIV-negative points.

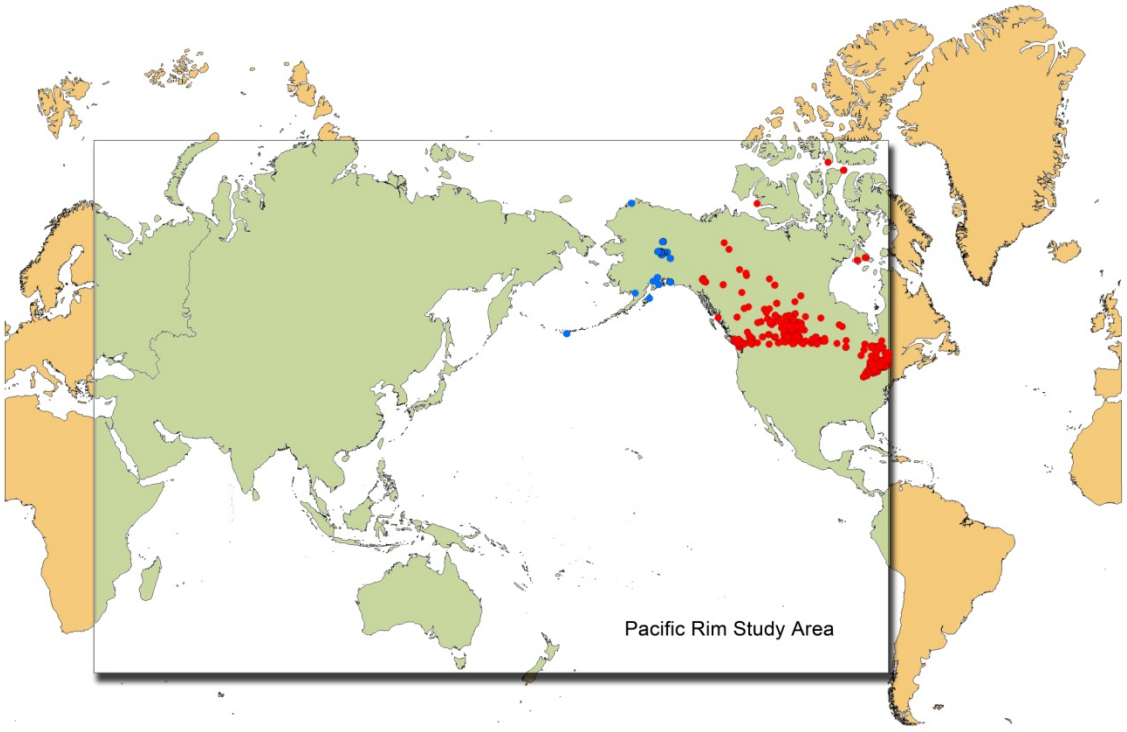

Figure 1.3. Pacific Rim study area and wild bird surveillance locations. These data are from the datasets used in Chapter 4. The bounding rectangle of the study area is represented by the green region within the box. The points indicate where sampling for avian influenza virus (AIV) occurred and include both AIV-positive and AIV-negative points. Sample locations from the Alaska Asia Avian Influenza Research database (2005-2010) are represented by the blue dots. Sample locations from the Canada's Inter-agency Wild Bird Influenza survey are represented by red dots.

LITERATURE CITED

[1] Krauss S, Walker D, Pryor P, Niles L, Chenghong L, Hinshaw VS, Webster RG: **Influenza A viruses of migrating wild aquatic birds in North America** *Vector Borne Zoonotic Dis* 2004, **4**:177-189.

[2] Munster VJ, Baas C, Lexmond P, Waldenstrom J, Wallensten A, Fransson T: **Spatial, temporal, and species variation in prevalence of influenza A viruses in wild migratory birds** *PLoS Path* 2007, **3**:630-638.

[3] Olsen B, Munster VJ, Wallensten A, Waldenström J, Osterhaus ADME, Fouchier RAM: **Global patterns of influenza A virus in wild birds** *Science* 2006, **312**:384-388.

[4] Hatta M, Gao P, Halfmann P, Kawaoka Y: **Molecular basis for high virulence of Hong Kong H5N1 influenza A viruses** *Science* 2001, **293**:1840-1842.

[5] World Organization for Animal Health (OIE): **Chapter 2.3.4 Avian Influenza.** In *Manual of Diagnostic Tests and Vaccines for Terrestrial Animals 2012*. 7th edition. Paris: OIE; 2005:Online edition [http://www.oie.int/fileadmin/Home/eng/Health_standards/tahm/2.03.04_AI.pdf].

[6] Li Z, Jiang Y, Jiao P, Wang A, Zhao F, Tian G, Wang X: **The NS1 gene contributes to the virulence of H5N1 avian influenza viruses** *J Virol* 2006, **80**:11115-11123.

[7] Fouchier RAM, Munster VJ: **Epidemiology of low pathogenic avian influenza viruses in wild birds** *Revue Scientifique et Technique - Office International des Épizooties* 2009, **28**:49-58.

[8] Garten RJ, Davis CT, Russell CA, Shu B, Lindstrom S, Balish A: **Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans** *Science* 2009, **325**:197-201.

[9] Ungchusak K, Auewarakul P, Dowell SF, Kitphati R, Auwanit W, Puthavathana P: **Probable person-to-person transmission of avian influenza A (H5N1)** *N Engl J Med* 2005, **352**:333-340.

[10] Butt KM, Smith GJD, Chen H, Zhang LJ, Leung YHC, Xu KM, Lim W, Webster RG: **Human infection with an avian H9N2 influenza A virus in Hong Kong in 2003** *J Clin Microbiol* 2005, **43**:5760-5767.

[11] Kandun IN, Wibisono H, Sedyaningsih ER, Yusharmen DPH, Hadisoedarsuno W, Purba W: **Three Indonesian clusters of H5N1 virus infection in 2005** *N Engl J Med* 2006, **355**:2186-2194.

[12] Subbarao K, Klimov A, Katz J, Regnery H, Lim W, Hall H: **Characterization of an avian influenza A (H5N1) virus isolated from a child with a fatal respiratory illness** *Science* 1998, **279**:393-396.

[13] Tran T, Nguyen T, Nguyen T, Luong T, Pham P, Nguyen V, Pham T: **Avian influenza A (H5N1) in 10 patients in Vietnam** *N Engl J Med* 2004, **350**:1179-1188.

[14] World Health Organization: Cumulative number of confirmed human cases for avian influenza A (H5N1) reported to WHO, 2003-2012 [http://www.who.int/influenza/human_animal_interface/EN_GIP_20120810CumulativeNumberH5N1cases.pdf].

[15] Bridges CB, Lim W, Hu-Primmer J, Sims L, Fukuda K, Mak KH, Rowe T, Katz JM: **Risk of influenza Z (H5N1) infection among poultry workers, Hong Kong** *J Infect Dis* 2002, **185**:1005-1010.

[16] Barbour EK, Sagherian VK, Sagherian NK, Dankar SK, Jaber LS, Usayran NN, Farran MT: **Avian influenza outbreak in poultry in the Lebanon and transmission to neighbouring farmers and swine** *Vet Ital* 2006, **42**:77-85.

[17] Lignon BL: **Avian influenza virus H5N1: a review of its history and information regarding its potential to cause the next pandemic** *Semin Pediatr Infect Dis* 2005, **16**:26-35.

[18] Webster R: **Predictions for future human influenza pandemics** *J Infect Dis* 1997, **176**:S14-19.

[19] Peiris M, Yuen K, Leung C, Chan K, Ip P, Lai R, Orr W, Shortridge K: **Human infection with influenza H9N2** *Lancet* 1999, **354**:916-917.

[20] Gill JS, Webby R, Gilchrist MJR, Gray GC: **Avian influenza among waterfowl hunters and wildlife professionals** *Emerg Infect Dis* 2006, **12**:1284-1286.

[21] World Health Organization: Regional Office for Africa: *Influenza pandemic risk assessment and preparedness in Africa*. Brazzaville; 2005.

[22] Chan M: World Health Organization: *World now at the start of 2009 influenza pandemic*. 2009.

[23] Fuller TL, Saatchi SS, Curd EE, Toffelmeier E, Thomassen HA, Buermann W, Smith TB: **Mapping the risk of avian influenza in wild birds in the U.S.** *BMC Infect Dis* 2010, **10**:187.

[24] Kaleta E, Pena K, Yilmaz A, Redmann T, Hofheinz S: **Avian influenza A viruses in birds of the order Psittaciformes: reports on virus isolations, transmission experiments and vaccinations and initial studies on innocuity and efficacy of oseltamivir in ovo** *Dtsch Tierarztl Wochenschr* 2007, **114**:260-267.

[25] Hall JS, Ip HS, Franson JC, Meteyer C, Nashold S, TeSlaa JL: **Experimental infection of a North American raptor, American kestrel (*Falco sparverius*), with highly pathogenic avian influenza virus (H5N1)** *PLoS ONE* 2009, **4**:e7555.

[26] Abolnik C, Gerdes G, Sinclair M, Ganzevoort B, Kitching J, Burger C: **Phylogenetic analysis of influenza A viruses (H6N8, H1N8, H4N2, H9N2, H10N7) isolated from wild birds, ducks, and ostriches in South Africa from 2007 to 2009** *Avian Dis* 2010, **54**:313-322.

[27] National Wildlife Health Center: List of species affected by H5N1 (avian influenza) [http://www.nwhc.usgs.gov/disease_information/avian_influenza/affected_species_chart.jsp].

[28] Kuiken T, Rimmelzwaan G, vanRiel D, vanAmerongen G, Baars M, Fouchier R, Osterhaus A: **Avian H5N1 influenza in cats** *Science* 2004, **306**:241.

[29] Songserm T, Amonsin A, Jam-on R, Sae-Heng N, Pariyothorn N, Payungporn S, Theamboonlers A: **Fatal avian influenza A H5N1 in a dog** *Emerg Infect Dis* 2006, **12**:1744-1747.

[30] Lipatov lS, Kwon YK, Sarmento LV, Lager KM, Spackman E, Suarez DL, Swayne DE: **Domestic pigs have low susceptibility to H5N1 highly pathogenic avian influenza viruses** *PLoS Path* 2008, **4**:e1000102.

[31] Gillim-Ross L, Santos C, Chen Z, Aspelund A, Yang C-F, Ye D, Jin H, Kemble G, Subbarao K: **Avian influenza h6 viruses productively infect and cause illness in mice and ferrets** *J Virol* 2008, **82**:10854-10863.

[32] Keawcharoen J, Oraveerakul K, Kuiken T, Fouchier RAM, Amonsi A, Payungporn S, Noppornpanth S: **Avian influenza H5N1 in tigers and leopards** *Emerg Infect Dis* 2004, **10**:2189-2191.

[33] Klopfleischa R, Wolf PU, Wolf C, Hardera T, Staricka E, Niebuhra M, Mettenleitera TC, Teifk JP: **Encephalitis in a stone marten (*Martes foina*) after natural infection with highly pathogenic avian influenza virus subtype H5N1** *J Comp Pathol* 2007, **137**:155-159.

[34] Anthony SJ, Leger JAS, Pugliares K, Ip HS, Chan JM, Carpenter ZW, Lipkin WI: **Emergence of fatal avian influenza in New England harbor seals** *mBio* 2012, **3**:e00166-00112.

[35] Shriner SA, VanDalen KK, Mooers NL, Ellis JW, Sullivan HJ, Root JJ, Pelzel AM, Franklin AB: **Low-pathogenic avian influenza viruses in wild house mice** *PLoS ONE* 2012, **7**:e39206.

[36] Shortridge KF, Gaob P, Guan Y, Ito T, Kawaoka Y, Markwella D, Takada A, Webster RG: **Interspecies transmission of influenza viruses: H5N1 virus and a Hong Kong SAR perspective** *Vet Microbiol* 2000, **74**:141-147.

[37] Perkins L, Swayne D: **Pathobiology of A/Chicken/Hong Kong/220/97 (H5N1) avian influenza virus in seven gallinaceous species** *Vet Pathol* 2001, **38**:149-164.

[38] Suarez D: **Evolution of avian influenza viruses** *Vet Microbiol* 2000, **74**:15-27.

[39] Boonyanuwat K, Thummabutra S, Sookmanee N, Vatchavalkyu V, Siripholvat V: **Influences of major histocompatibility complex class I haplotypes on avian influenza virus disease traits in Thai indigenous chickens** *Anim Sci J* 2006, **77**:285-289.

[40] Boonyanuwat K, Thummabutra S, Sookmanee N, Vatchavalkhu V, Siripholvat V, Mitsuhashi T: **Influences of MHC class II haplotypes on avian influenza traits in Thai indigenous chickens** *Poult Sci* 2006, **43**:120-125.

[41] Berthouly-Salazar C, Rognon X, Van TN, Gély M, Chi CV, Tixier-Boichard M, Bed'Hom B, Bruneau N, Verrier E, Maillard J, Michaux J: **Vietnamese chickens: a gate towards Asian genetic diversity** *BMC Genetics* 2010, **11**:53.

[42] Granevitze Z, Hillel J, Chen GH, Cuc NTK, Feldman M, Eding H, Weigend S: **Genetic diversity within chicken populations from different continents and management histories** *Anim Genet* 2007, **38**:576-583.

[43] Doeschl-Wilson AB, Brindle W, Emmans G, Kyriazakis I: **Unravelling the relationship between animal growth and immune response during micro-parasitic infections** *PLoS ONE* 2009, **4**:e7508.

[44] Kim J-K, Negovetich NJ, Forrest HL, Webster RG: **Ducks: The "Trojan Horses" of H5N1 influenza** *Influenza Other Respi Viruses* 2009, **3**:121-128.

[45] Hulse-Post D, Sturm-Ramirez K, Humberd J, Seiler P, Govorkova EA, Krauss S, Scholtissek C: **Role of domestic ducks in the propagation and biological evolution of highly pathogenic H5N1 influenza viruses in Asia** *Proc Natl Acad Sci USA* 2005, **102**:10682-10687.

[46] Halvorson DA, Kelleher CJ, Senne DA: **Epizootiology of avian influenza: effect of season on incidence in sentinel ducks and domestic turkeys in Minnesota** *Appl Environ Microbiol* 1985, **49**:914-919.

[47] Halvorson D, Capua I, Cardona C, Frame D, Karunakaran D, Marangon S, Ortali G, Roepke D, Woo-Ming B: **The economics of avian influenza control**. In: *Proceedings of the 52nd Western Poultry Disease Conference: 8-11 March 2003; Sacramento.* 2003: 5-7.

[48] FAO Emergency Prevention System (EMPRES)/Global Early Warning System (GLEWS): Agriculture Department/Animal Production and Health Division: *H5N1 HPAI global overview: January-March 2012*. 2012.

[49] Cutler DR, Edwards TC, Jr., Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ: **Random Forests for classification in ecology** *Ecology* 2007, **88**:2783-2792.

[50] USDA Animal and Plant Health Inspection Service: Avian Influenza Surveillance in Meat-Type Chickens (Broilers) [http://www.aphis.usda.gov/vs/nahss/poultry/ai/index.htm].

[51] Li J, Dohna Hz, Anchell NL, Adams SC, Dao NT, Xing Z, Cardona CJ: **Adaptation and transmission of a duck-origin avian influenza virus in poultry species** *Virus Res* 2010, **147**:40-46.

[52] Lee H, Kwon J, Lee D, Lee Y, Youn H, Kim M: **Continuing evolution and interspecies transmission of influenza viruses in live bird markets in Korea** *Avian Dis* 2010, **54**:738-748.

[53] Li J, Cardona CJ: **Adaptation and transmission of a wild duck avian influenza isolate in chickens** *Avian Dis* 2010, **54**:586-590.

[54] Ito T, Goto H, Yamamoto E, Tanaka H, Takeuchi M, Kuwayama M, Kawaoka Y, Otsuki K: **Generation of a highly pathogenic avian influenza A virus from an avirulent field isolate by passaging in chickens** *J Virol* 2001, **75**:4439-4443.

[55] Ramirez-Nieto G, Shivaprasad HL, Kim C-H, Lillehoj HS, Song H, Osorio IG, Perez DR: **Adaptation of a mallard H5N2 low pathogenicity influenza virus in chickens with prior history of infection with infectious bursal disease virus** *Avian Dis* 2010, **54**:513-521.

[56] Cilloni F, Toffan A, Giannecchini S, Clausi V, Azzi A, Capua I, Terregino C: **Increased pathogenicity and shedding in chickens of a wild bird-origin low pathogenicity avian influenza virus of the H7N3 subtype following multiple in vivo passages in quail and turkey** *Avian Dis* 2010, **54**:555-557.

[57] Soda K, Asakura S, Okamatsu M, Sakoda Y, Kida H: **H9N2 influenza virus acquires intravenous pathogenicity on the introduction of a pair of di-basic amino acid residues at the cleavage site of the hemagglutinin and consecutive passages in chickens** *Virol J* 2011, **8**:64.

[58] Hinshaw VS, Sheerar MG, Larsen D: **Specific antibody responses and generation of antigenic variants in chickens immunized against a virulent avian influenza virus** *Avian Dis* 1990, **34**:80-86.

[59] Shengqing Y, Kishida N, Ito H, Kida H, Otsuki K, Kawaoka Y, Ito T: **Generation of velogenic Newcastle disease viruses from a nonpathogenic waterfowl isolate by passaging in chickens** *Virology* 2002, **301**:206-211.

[60] Slemons RD, Easterday BC: **Virus replication in the digestive tract of ducks exposed by aerosol to type-A influenza** *Avian Dis* 1978, **22**:367-377.

[61] Kida H, Yanagawa R, Matuoka Y: **Duck influenza lacking evidence of disease signs and immune response** *Infect Immun* 1980, **30**:547-553.

[62] Jourdain E, Gunnarsson G, Wahlgren J, Latorre-Margalef N, Bröjer C, Sahlin S, Svensson L, Waldenström J, Lundkvist Å, Olsen B: **Influenza virus in a natural host, the mallard: experimental infection data** *PLoS ONE* 2010, **5**:e8935.

[63] Capua I, Mutinelli F: **Mortality in Muscovy ducks (*Cairina moschata*) and domestic geese (*Anser anser var. domestica*) associated with natural infection with a highly pathogenic avian influenza virus of H7N1 subtype** *Avian Pathol* 2001, **30**:179-183.

[64] Browning DM, Beaupre SJ, Duncan L: **Using Partitioned Mahalanobis D2(k) to Formulate a GIS-Based Model of Timber Rattlesnake Hibernacula** *J Wildl Manage* 2005, **69**:33-44.

[65] Sturm-Ramirez KM, Ellis T, Bousfield B, Bissett L, Dyrting K, Rehg JE, Poon L, Guan Y: **Reemerging H5N1 influenza viruses in Hong Kong in 2002 are highly pathogenic to ducks** *J Virol* 2004, **78**:4892-4901.

[66] van Gils JA, Munster VJ, Radersma R, Liefhebber D, Fouchier RAM, Klaassen M: **Hampered Foraging and Migratory Performance in Swans Infected with Low-Pathogenic Avian Influenza A Virus** *PLoS ONE* 2007, **2**:e184.

[67] Latorre-Margalef N, Gunnarsson G, Munster VJ, Fouchier RAM, Osterhaus ADME, Elmberg J, Olsen Br: **Effects of influenza A virus infection on migrating mallard ducks** *Proc R Soc* 2009, **275**:1029-1036.

[68] Peterson AT, Sanchez-Cordero V, Beard CB, Ramsey JM: **Ecological niche modeling and potential reservoirs for Chagas disease, Mexico** *Emerg Infect Dis* 2002, **8**:662-667.

[69] Furlanello C, Neteler M, Merler S, Menegon S, Fontanari S, Donini A, Rizzoli A, Chemini C: **GIS and the Random Forest predictor: integration in R for tick-borne disease risk assessment**. In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing: 20-22 March 2003; Vienna.* Edited by Hornik K, Leisch F, Zeileis A; 2003: 1-11.

[70] Mak S, Morshed M, Henry B: **Ecological niche modeling of Lyme Disease in British Columbia, Canada** *J Med Entomol* 2010, **47**:99-105.

[71] Moffett A, Shackelford N, Sarkar S: **Malaria in Africa: vector species' niche models and relative risk maps** *PLoS ONE* 2007, **2**:e824.

[72] Peterson AT, Vieglais DA, Andreasen JK: **Migratory birds modeled as critical transport agents for West Nile Virus in North America** *Vector Borne Zoonotic Dis* 2003, **3**:27-37.

[73] Guerra M, Walker E, Jones C, Paskewitz S, Cortinas MR, Stancil A, Beck L, Bobo M, Kitron U: **Predicting the risk of Lyme disease: habitat suitability fo *Ixodes scapularis* in the North Central United States** *Emerg Infect Dis* 2002, **8**:289-297.

[74] Glass G, Schwartz B, Morgan J, Johnson D, Noy P, Israel E: **Environmental risk factors for Lyme disease identified with geographic information systems** *Am J Public Health Nations Health* 1995, **85**:944-948.

[75] Brownstein JS, Holford TR, Fish D: **Effect of climate change on Lyme disease risk in North America** *EcoHealth* 2005, **2**:38-46.

[76] Tran A, Goutard F, Chamaille L, Baghdadi N, Seen DL: **Remote sensing and avian influenza: A review of image processing methods for extracting key variables affecting avian influenza virus survival in water from Earth Observation satellites** *Int J Appl Earth Obs Geoinf* 2010, **12**:1-8.

[77] Clark JD, Dunn JE, Smith KG: **A multivariate model of female black bear habitat use for a geographic information system** *J Wildl Manage* 1993, **57**:519-526.

[78] Kilpatrick AM, Chmura AA, Gibbons DW, Fleischer RC, Marra PP, Daszak P: **Predicting the global spread of H5N1 avian influenza** *Proc Natl Acad Sci USA* 2006, **103**:19368-19373.

[79] Adhikari D, Chettri A, Barik SK: **Modelling the ecology and distribution of highly pathogenic avian influenza (H5N1) in the Indian subcontinent** *Curr Sci* 2009, **97**:72-78.

[80] Williams RAJ, Fasina FO, Peterson AT: **Predictable ecology and geography of avian influenza (H5N1) transmission in Nigeria and West Africa** *Trans R Soc Trop Med Hyg* 2008, **102**:471-479.

[81] Gilbert M, Chaitaweesub P, Parakamawongsa T, Premashthira S, Tiensin T, Kalpravidh W, Wagner H, Slingenbergh J: **Free-grazing ducks and highly pathogenic avian influenza, Thailand** *Emerg Infect Dis* 2006, **12**:227-234.

[82] Gilbert M, Xiao X, Pfeiffer DU, Spprecht M, Boles S, Czarnecki C: **Mapping H5N1 highly pathogenic avian influenza risk in Southeast Asia** *Proc Natl Acad Sci USA* 2008, **105**:4769-4774.

[83] Pfeiffer DU, Minh PQ, Martin V, Epprecht M, Otte MJ: **An analysis of the spatial and temporal patterns of highly pathogenic avian influenza occurrence in Vietnam using national surveillance data** *Vet J* 2007, **174**:302-309.

[84] Peterson AT: **Ecologic niche modeling and spatial patterns of disease transmission** *Emerg Infect Dis* 2006, **12**:1822-1826.

[85] Souris M, Gonzalez J-P, Shanmugasundaram J, Corvest V, Kittayapong P: **Retrospective space-time analysis of H5N1 Avian Influenza emergence in Thailand** *Int J Health Geogr* 2010, **9**.

[86] Reemers SS, Leenen Dv, Koerkamp MJG, Haarlem Dv, Haar Pvd, Eden Wv, Vervelde L: **Early host responses to avian influenza A virus are prolonged**

and enhanced at transcriptional level depending on maturation of the immune system *Mol Immunol* 2010, **47**:1675-1685.

[87] Hogerwerf L, Walker RG, Ottaviani D, Slingenbergh J, Prosser D, Bergmann L, Gilbert M: **Persistence of highly pathogenic avian influenza H5N1 virus defined by agro-ecological niche** *EcoHealth* 2010, **7**:213-225.

[88] Breiman L: **Random Forests** *Mach Learn* 2001, **45**:5-32.

[89] Herrick KA, Huettmann F, Runstadler J, Chernetsov N, Antonov A, Valchuk O, Gerasimov Y, Matsyna E, Matsyna A, Markovets M, Druzyaka A, Saito K: **Predictive RISK modeling of avian influenza in the Pacific Rim and beyond.** In *Risk Models and Applications, 2010*. Edited by Kremers H, Susini A. Berlin: CODATA Germany: Lecture Notes in Information Sciences; 2010:135-148.

CHAPTER 2

Predictive risk modeling of avian influenza around the Pacific Rim[1]

ABSTRACT

In the past 10 years, outbreaks of highly pathogenic avian influenza (HPAI) have raised concerns over the potential for zoonotic influenza pandemic. Quantitative predictions of risk can be determined using data from ongoing influenza surveillance efforts combined with advanced statistical methods. In particular, we were interested in modeling the risk posed by wild bird populations, which are the reservoir of avian influenza virus (AIV) and maintain a pool of far greater viral diversity than that found in domestic poultry. We chose to focus our study area on the Pacific Rim as it encompasses three major migratory flyways and the Alaska summer breeding grounds where they all overlap. The purpose of this study was to develop a premier predictive niche model based on field data, laboratory analysis, and Geographic Information Systems (GIS) data and map the predicted relative occurrence of AIV in wild birds. Using bioclimatic data layers, we applied the ensemble data-mining algorithm Random Forests to form a landscape-scale prediction of where AIV is expected to occur in wild birds. Important predictors correlated with AIV-

[1] Originally published as
Herrick, K.A., F. Huettmann, J.A. Runstadler, et al. 2010. Predictive RISK modeling of avian influenza in the Pacific Rim and beyond, In: Kremers, H., Susini, A. (Eds.), Risk Models and Applications, 2010. CODATA Germany: Lecture Notes in Information Sciences, Berlin, p. 190.
The same data used in this publication underwent additional grooming and were re-analyzed. The paper was re-written for clarity and the authors were reduced to those who contributed directly to the research and manuscript.

positivity included low temperatures in February (-20 °C), low temperatures in November (-10.5 °C), high temperature seasonality (140 °C), and a long distance from the coast (800km). We described the niche of AIV-positive cases as inland regions with a continental climate and very cold winter temperatures. This model will provide a collaborative digital and empirical framework into which other researchers in the avian influenza community can contribute and share information for further improvements. As of this writing, this project is one of the first of its kind representing a predictive map of low-pathogenicity avian influenza in wild birds.

INTRODUCTION

The threat of a human "bird flu" pandemic focused a great deal of attention on a highly pathogenic Eurasian strain of the H5N1 avian influenza virus (AIV) subtype. This strain continues to create economic losses [1] and demonstrate a threat to human health [2, 3]. However, when one considers that most of the 170 possible combinations of hemagglutinin (H) and neuraminidase (N) protein subtypes have been isolated from wild birds [4-6], it is clear that HPAI H5N1 cases represent a very small subset within the genetic diversity of AIV. Most viral strains are classified as low-pathogenicity avian influenza (LPAI), based on their lethality to chickens [7], and rare strains of H5 and H7 subtypes are highly pathogenic. AIV is endemic in populations of wild birds, and waterfowl (Anseriformes) are the most commonly infected with AIV, followed by shorebirds and gulls (Charadriiformes), and perching birds (Passeriformes) [8]. Infection in other birds including Psittaciformes [9], raptors [10], shorebirds [11], and ratites [12] have also been documented.

Mallards (*Anas platyrhynchos*) infected with virus are typically asymptomatic [13-15], which may enable them to act as highly mobile carriers of influenza virus, especially when migrating. It is not uncommon for passing wild birds, attracted by water and food, to intermingle with domestic poultry and swine, exposing them to the virus. A number of documented HPAI cases in domestic poultry have been traced back to wild birds [12, 16-19], including outbreaks of H6 and H9 subtypes, which are normally considered LPAI. Wild waterfowl are a greater source of outbreaks than endemic AIV strains already circulating in domestic poultry [20]. Clearly, AIV risk management

strategies must consider all viral subtypes, not just H5N1, and include wild birds rather than limiting studies to domestic poultry.

Concerns over "bird flu" increased the interest in AIV surveillance projects, which report the viral subtypes isolated each year. Since 2005, an extensive AIV surveillance project has been conducted by the Alaska Asia Avian Influenza Research group (A3IR), a sub-group within the NIH-NIAD Centers for Excellence in Influenza Research and Surveillance (CEIRS). The overall goal of the CEIRS program is to provide decision-makers, scientists, and the global audience with information, tools, and strategies needed to "control and lessen the impact of epidemic influenza and the increasing threat of pandemic influenza" [21]. Biologists, veterinarians, and laboratory virologists make up A3IR, which includes Russian, Japanese, US, and Mongolian collaborators. The study area encompasses the Mississippi Americas flyway, Pacific Americas flyway, and the East Asia/Australian flyway. All three flyways overlap in Alaskan summer breeding grounds, where North and South American populations potentially come in contact with Asian populations. While the HPAI H5N1 subtype has not yet been identified in the Americas, other Asian strains of AIV have been isolated from birds sampled in Alaska [22, 23].

Availability of A3IR surveillance data allowed us to use data-mining and statistical techniques to begin quantifying the risk posed by wild waterfowl within this important study area. When combined with advanced statistical prediction algorithms, GIS has the power to integrate diverse types of data in order to describe and detect complex patterns, then quantitatively model and develop predictions. Remote sensing

data and Geographic Information Systems (GIS) technology have proven useful in the management of avian influenza. A number of studies (reviewed in Tran et al. [24]) have used remote sensing data to identify individual environmental indicators (e.g. wind, frost, rice cultivation [17]) related to the spread of AIV. Ecological niche modeling has more commonly been used in parasite-host cases such as Chagas disease [25], Lyme disease [26], malaria [27], and West Nile virus [28]. In the case of a disease like AIV, niche modeling can supplement field and lab work by predicting prevalence in areas that are otherwise difficult to survey, assist in the development of hypotheses, and guide collection by characterizing areas where AIV may be found. Considering the important sampling range of A3IR, an excellent opportunity exists to provide the CEIRS project with information beyond descriptive statistics.

The purpose of this project is to create a model describing the environmental niche in which AIV-positive samples are collected and use this information to predict AIV-positive locations beyond sampled locations. This predictive model will help us to better understand and test the effects of environmental and spatial patterns on the dynamics of ecological processes influencing AIV persistence on a landscape scale. We used the ensemble classification-tree algorithm Random Forests to identify important bioclimatic and anthropogenic conditions that contribute to the occurrence of AIV-positivity. Based on these factors, we created a predictive map of the relative occurrence of avian influenza across the study area of the Pacific Rim. Our study included in its analysis over 20,000 points of wild bird data collected by this international team between 2005 and 2007. It is the first of its kind to use a study area of this size encompassing the

Pacific Rim, and it is unique because it includes all available viral subtypes rather than limiting its scope to HPAI H5N1.

MATERIALS AND METHODS

Data layers

In order to define the niche of AIV positive locations, we compiled three types of predictors: bird, bioclimatic, and geographic data layers. Bird data were geo-referenced samples of cloacal, oropharyngeal, and fecal swabs collected from wild birds by A3IR collaborators as previously described [29]. Collectors recorded an extensive set of attributes for each bird such as the GPS coordinates of collection location, species, approximate age, health condition of the bird, date of collection, and description of location. As we were interested in examining bioclimatic correlates of AIV-positivity, attributes such as bird biometrics were excluded from analysis. The AIV-positive status of each individual sample was determined by collaborating labs. The original sample swabs or cDNA amplified from the swabs were sent to the University of Alaska Fairbanks for analysis and viral subtype identification by RRT-PCR [30]. The Russian samples, which could not be sent to the US, were analyzed at VECTOR State Research Center of Virology and Biotechnology in Novosibirsk, Russia. Japanese samples were analyzed at the Obihiro University of Agriculture and Veterinary Medicine. Samples were considered AIV-positive if viral matrix screening indicated the presence of virus. Bird data and analysis results were entered into the main database at the University of Alaska Fairbanks (F. Aldehoff and R. Koskela, *pers. comm.*). As the bird data had been

collected and entered by many users and entries had not been curated, we groomed the data prior to analysis. Incongruities between GPS coordinates and location description were corrected if possible; irreconcilable points were excluded from analysis. Entries were excluded from analysis if they were not tested for AIV, if the bird species was listed as "Unknown", or if the sample type was listed as "environmental" as the species could not be positively identified. Calculations were performed to make data consistent in format, such as converting GPS coordinates from degrees to decimal degrees. Additional binary columns were created to indicate states such as AIV-positive status and approximate age (hatch-year juveniles and post hatch-year adults). The final database included 189 different bird species, and contained 21,149 data points with an AIV prevalence of 3.5%. A list of bird species including the total number of samples collected and the number that tested positive for AIV are presented as Appendix E.

Geographic and bioclimatic predictor variable data layers were acquired as high-quality global data layers from open access projects. Bioclimatic variables useful for defining niche conditions, such as annual mean temperature and monthly precipitation were obtained from WorldClim [31]. As waterfowl make up a large proportion of our samples, we included a global hydrology layer [32] and calculated the distance from rivers, lakes, and wetlands using the Euclidean distance tool in ArcGIS. *Charadriiformes* such as gulls and wading birds are another important reservoir of AIV, so we calculated Euclidean distance from coastline. As human presence and activity may impact bird health, e.g. adverse effects from pollution or improved forage in agricultural regions, we included layers reflecting indices of human influence obtained from Last of the Wild,

version 2 [33]. The Human Influence and Human Footprint indices are calculated based on human population density, land transformation, transportation infrastructure, and electrical power infrastructure. Human population density was obtained from the Center for International Earth Science Information Network (CIESIN) [34]. In order to integrate risk factors associated with agricultural livestock, we included layers for predicted pig and poultry densities obtained from FAO GeoNetwork global livestock distributions [35]. All predictor variable layers are listed in Table 2.1.

Modeling methods

In order to determine the predictor variables with the highest contribution to characterizing the environmental niche for AIV, we employed the Random Forests algorithm implemented in Salford Predictive Miner (Salford Systems). Random Forests is an ensemble classification (and regression) tree method, which has demonstrated robust predictive power in ecological settings [36]. This algorithm differs from other classification tree methods in that the best split at each node is selected from a small subset of predictor variables, which are randomly drawn at each node. Each tree is then grown to its fullest extent without pruning. The process is repeated multiple times to produce a "forest" of trees. The strength and accuracy of the model is sensitive to the number of predictor variables at each node. However, the optimal range is wide and easily predicted; the value that is typically used is the square root of the total number of predictors. Random Forests is considered relatively immune to overfitting, or construction of an overly complex model that describes noise in the data rather than describing the relationship between predictors that produce the signal [37]. In addition,

Random Forests calculates its own internal, unbiased running estimate of the model's classification error. For each tree, approximately two thirds of the total number of cases are used to grow the tree, while the remaining third are kept "out-of-bag" (OOB) for error testing. Once the tree has been constructed, the OOB samples are run down the tree and OOB error is calculated as the proportion of times these samples are misclassified, averaged over all cases. The Random Forests algorithm also calculates the contribution of each predictor variable to the accuracy of the model in two different ways: raw importance and Mean Decrease in Gini. The raw importance indicates the contribution of a variable to model accuracy over that of a random variable. Raw importance is calculated by permuting the values of a variable (making it a random, unrelated variable) and running it down a tree. The correctly categorized cases from the permuted set are subtracted from the correctly categorized cases from the OOB set, and this value is averaged across the forest to produce a raw importance score. 'Mean Decrease' (in the Gini setting) is calculated by summing the decrease in node impurity for each variable over the entire forest. Random Forests then produces ranked lists of variables based on these measures of importance.

Using Random Forest, we were able to predict areas of AIV positivity. Briefly, we trained the model on 52 predictors (Table 1) with AIV-positivity as the target variable. We then applied the resulting model to a regularly-spaced grid of 50,000 points that covered the study area, and predicted the AIV-positivity of each point. These results were interpolated across the study area using the Inverse Distance Weighted (IDW) function in ArcGIS producing a map of the Relative Occurrence Index of AIV.

Model evaluation

We evaluated the prediction accuracy of this model using the Receiver Operating Characteristic (ROC) summary statistic, which is based on a confusion matrix of correctly and incorrectly predicted positive and negative points. This method plots the true positive rate against false positive rate and calculates the area under the resulting curve (AUC). Rather than using p-values and significance, we chose ROC as the most suitable metric for assessing predictions in a large and complex landscape context [38]. Random Forests also produced a ranked list of predictor variables in order of their Mean Decrease in Accuracy score, which indicates their contribution to model accuracy. To compare AIV-positive and –negative cases, we further characterized the top four predictors, which received Importance Scores over 50, using notched box plots and histogram density plots generated by S+ analytical software (TIBCO Spotfire, v.8.2). In order to summarize the individual effects of each predictor above the averaged effects of the other predictors, we generated partial dependence plots using the randomForest package [39] in R Statistical Programming Language [40].

RESULTS

Random Forests produced a robust model allowing us to visualize and map the predicted relative occurrence of avian influenza in wild birds and characterize the environmental niche in which it is predicted to occur. In Figure 2.1 we present the resulting map of the predicted relative occurrence index where high risk areas are red and low risk areas are blue. This model had an ROC of 0.841 demonstrating a high level of model performance.

Mean temperature in February contributed the most to model accuracy followed by mean temperature in November, temperature seasonality (variability in seasonal temperature calculated as the °C standard deviation of mean monthly temperatures x 100), and Euclidean distance from coastline (Table 2.2). The mean temperature in February for AIV-positive and AIV–negative cases have similar medians (approximately -20 °C; Fig. 2.2A) and means (-19.5 °C and -16.9 °C, respectively). The range over which the AIV-negative samples were found is much broader than that of the AIV-positive samples (Fig. 2.3A), which display a strong peak around -20 °C. The partial dependence of AIV-positivity on the mean temperature in February is very high at low temperatures up to -20 °C, at which point it displays a strong threshold (Fig. 2.4A). AIV-positive cases occurred over a slightly higher temperature in November (mean of -10.5 °C) than AIV-negative cases (-15.4 °C). Similar to the mean temperature in February, the AIV-negative samples occurred over a much wider range than the AIV-positive cases (Fig. 2B, 3B). However, there are two density peaks for AIV-positive cases at approximately -17 °C and 5 °C (Fig. 2.3B) that are distinct from AIV-negative cases and the partial dependence on the mean temperature in November shows a very strong threshold at -15 °C. For both predictors, AIV-positivity appears to be strongly correlated with and dependent upon low temperatures. AIV-positive cases had a higher mean temperature seasonality than AIV-negative cases (140.2 °C and 121.5 °C, respectively), although their distributions and densities were similar. Both density and partial dependence show single strong peaks around this mean as well (Fig. 2.3C, Fig. 4C). Although the AIV-positive and AIV–negative cases appear to have nearly identical medians and distributions (Fig. 2.2D), the

mean distance from coastline for positive cases was much further inland (800 km) than negative cases (58 km). AIV-positive cases show a strong peak in density over AIV-negative cases at 800 km (Fig. 2.3D) and their partial dependence shows a strong threshold at this same value as well, changing from almost no dependence to very high dependence as distance from coast increases.

DISCUSSION

We are not presenting this model as an exhaustive prediction of AIV, but rather to present one of the first models of its kind, in that it 1) focused on LPAI, 2) used wild birds as subjects, and 3) was applied across the Pacific Rim, a much larger scale than in previous studies. This model presents a technical platform for applying data mining in an ecology and risk management context and demonstrates that these modeling techniques are robust enough to glean signal from noise in a system as complicated as avian influenza [41, 42]. Risk modeling approaches, such as those presented in this project, for HPAI [43] [44], and other diseases [45, 46], are necessary in the endeavor of minimizing the risk of AIV to humans and livestock and for managing human health concerns.

Due to the importance of wild birds in the risk management of AIV, we were interested in quantifying their contribution to the risk of AIV. Despite the complexity of an infectious viral disease system in migratory birds, we were able to characterize the environmental niche of AIV-positive cases, apply this model, and construct a predictive map around the Pacific Rim. This model predicts AIV at high latitude, inland regions, with cold winters, and a high variation in seasonal temperatures. Most of these areas

occur across northern North America and Siberia where a strong continental climate effect, and therefore a large variation in seasonal temperature, would be expected. Despite the high proportion of ducks and waterfowl that make up the sampled species, close proximity to lakes and rivers did not make a large contribution to model accuracy. Despite coastal sampling of gulls and wading birds, AIV-positivity appears to depend upon increasing distance from the coast. Hot spots show up on Tasmanian and New Zealand coasts, which is puzzling because February and November are their summer months and relatively stable climate is expected in these coastal regions. These interesting findings require testing in the field, and are of considerable relevance for temperature-related AIV questions, such as climate scenarios where increased temperature is predicted. Overall, the predicted risk to humans and their livestock is relatively low. Despite the fact that the flyways traverse areas of high population density, the AIV positive niche is primarily characterized as sparsely populated.

Our findings differed from Adhikari [47] who found a high human population density >100 persons/km^2 was an important predictor for avian influenza outbreak. Neither population density nor any other anthropogenic predictors made significant contributions to our model. The apparent contradiction highlights key differences between our study and theirs. First, their study incorporated bird data from the densely populated areas of West Bengal and Bangladesh interpolated across India, Bangladesh, Nepal, and part of Pakistan. Next, they focused on the HPAI sub-strain of H5N1 in domestic poultry. Human density and populated areas do not exclude wild birds, and can benefit bird species that are able to opportunistically inhabit niches created by

urbanization [48]. However, our study included only wild bird data collected in sparsely populated areas and viral subtypes only found in wild birds, which may partly explain why anthropogenic factors did not make a large contribution to our model. We expect that in a study on domestic poultry sampled in densely populated regions, human population density would tend to co-vary strongly with the species in question. Poultry density as a predictor variable in Adhikari model would be an interesting addition. This model is currently limited to A3IR data and we hope for improved global data sharing policies. Further collaboration would allow us to confront the model with more data, which would not only strengthen the prediction and improve the niche description, but also allow us to interpolate findings across a wider study area. As such, our model also poses an interesting function of risk modeling as a quality control step for surveillance data. Unexpected results can be examined as truly novel findings or as evidence of gaps in surveillance effort. The inclusion of lower latitude data would likely challenge our findings that northern regions are important in the predicted occurrence of AIV and extend the findings to more densely populated areas where AIV outbreak could pose serious health risk.

ACKNOWLEDGEMENTS

We greatly thank the following researchers (in alphabetical order) and their teams for their time and effort in the field collecting samples from 20,000+ birds: A. Antonov, S. Backensto, K. Beckman, T. Booms, N. Chernetsov, K. DeGroot, A. Dixon, A. Druzyaka, T. Eskelin, J. Fischer, Y. Gerasimov, R. Gerlach, L. Hawkins, D. Holcomb, K. Kloecker, A. Lang, M. Lindberg, A. Matsyna, E. Matsyna, M. Markovets, K. Martin, B. McCaffery, B. Meixell, T. Moran, J. Pearce, M. Petrula, A. Powell, D. Rosenberg, K. Saito, J. Saracco, J. Sedinger, S. Sharbough, A. Springer, A. Taylor, O. Valchuk, J. Whitman, K. Winker, and D. Zwiefelhofer. We thank the following laboratories and their technicians for analyzing field samples: UAF Core Lab, Jon Runstadler's Lab and George Happ's Lab at the University of Alaska Fairbanks, VECTOR (Novosibirsk), Richard Slemon's Lab, H. Ogawa and K. Imai from the Obihiro University of Agriculture and Veterinary Medicine, R. Halpin and D. Spiro at the J. Craig Venter Institute, J. Harris at the Los Alamos National Labs. We thank the Alaska Bird Observatory and the Alaska Raptor Center for sampling birds and sharing data. We thank G. Humphries and other members of the EWHALE lab at UAF for their invaluable expertise. Sampling effort and analysis was funded in part by the Alaska IDeA Networks of Biomedical Research Excellence (INBRE), part of NCRR's Division of Research Infrastructure Grant Number 5P20RR016466 and in part by Sub-Award Contract from NIAID Number HHSN266200700009C, a component of the National Institutes of Health (NIH). This project and its contents are solely the responsibility of the authors and do not necessarily represent the official views of NCRR, NIAID or NIH.Infrastructure Grant Number

5P20RR016466 and in part by Sub-Award Contract from NIAID Number HHSN266200700009C, a component of the National Institutes of Health (NIH). This project and its contents are solely the responsibility of the authors and do not necessarily represent the official views of NCRR, NIAID or NIH.

TABLES

Table 2.1. Predictor variables used to construct model of avian influenza in wild birds. Variables are listed here with details and source project.

Predictor Variable	Details	Project
Gridded population of the world, v.3	Persons/km^2	CIESIN [49]
Human Influence Index	Summative index of human disturbance (0-72)	Last of the Wild [33]
Human footprint index	Percentage of relative human influence (0-100)	"
Predicted global pig density	Animal density/km2 ; 3 min of arc	FAO: Gridded Livestock of the World [35]
Predicted global poultry density	"	"
Annual Mean Temperature	In C^o x 10; 30 arc-seconds, 1 km spatial resolution	WorldClim [31]
Annual Precipitation	In mm;	"
Elevation	In meters	"
Isothermality	In ^oC; 30 arc-seconds, 1 km spatial resolution	"
Max temp warmest month	In ^oC; 30 arc-seconds, 1 km spatial resolution	"
Mean diurnal range	Mean of monthly (max temp - min temp)	"
Mean temp, coldest quarter	In ^oC; 30 arc-seconds, 1 km spatial resolution	"
Mean temp, driest quarter	"	"
Mean temp, warmest quarter	"	"
Mean temp, wettest quarter	"	"
Mean temp, coldest month	"	"
Precipitation of coldest quarter	"	"
Precipitation of driest month	"	"
Precipitation of driest quarter	"	"
Precipitation of warmest quarter	"	"

Table 2.1 continued

Predictor Variable	Details	Project
Precipitation of wettest quarter	"	WorldClim [31]
Precipitation seasonality	Coefficient of variation	"
Temperature annual range	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	"
Temperature seasonality	Calculated as (standard deviation * 100)	"
Mean monthly temperature (January – December)	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	"
Mean monthly precipitation (January – December)	In mm; 30 arc-seconds, 1 km spatial resolution	"
Euclidean distance from coast	Calculated in ArcGIS	n/a
Euclidean distance from hydrologic feature	"	n/a

Table 2.2. Normalized importance scores for top predictor variables. Scores are based on the Mean Decrease in Accuracy values generated by Random Forests. Means for each predictor variable grouped by positivity or negativity for avian influenza virus (AIV) are also included.

Variable	Importance Score	AIV-positive mean	AIV-negative mean
Mean temperature in February (°C)	100	-19.5	-16.9
Mean temperature, November (°C)	69.0	-10.5	-15.4
Temperature seasonality (°C)	63.2	140.2	121.5
Distance from coast (km)	53.7	800	58

FIGURES

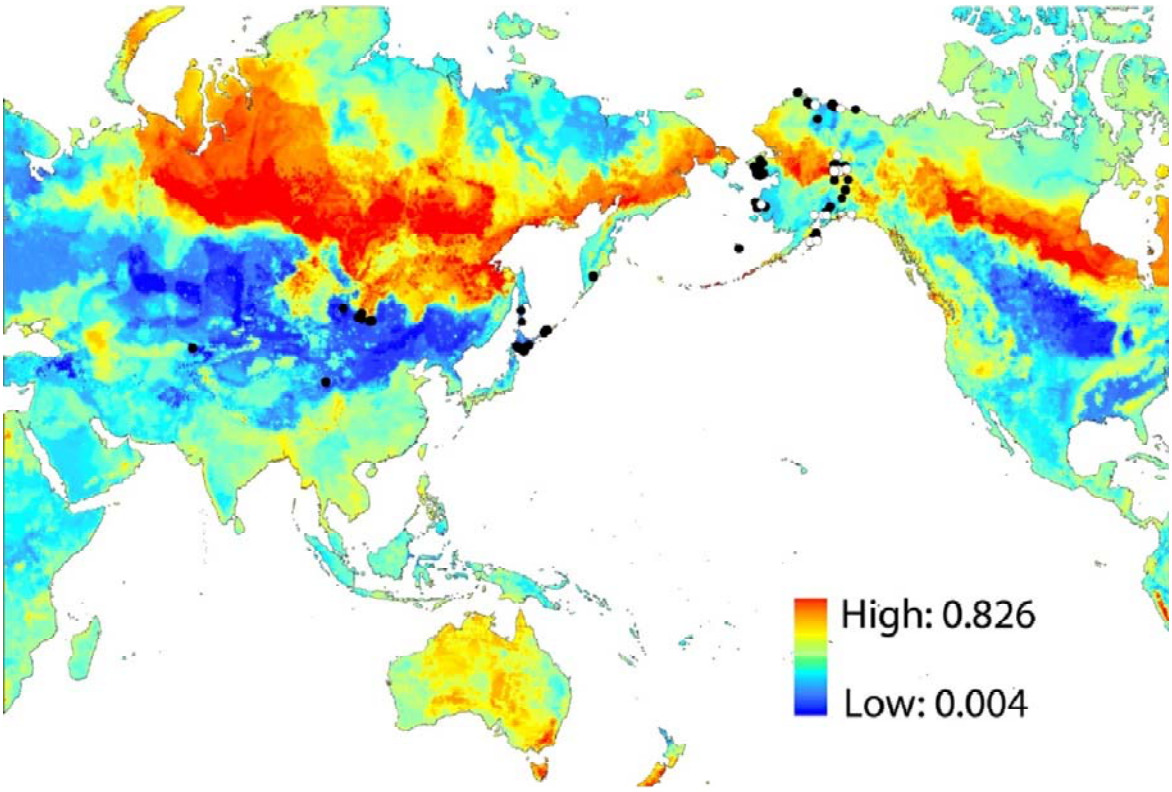

Figure 2.1. Map of predicted relative occurrence index of avian influenza virus (AIV) in wild birds around the Pacific Rim study area and surveillance locations. The map colors indicate areas of high (red) predicted occurrence of AIV to low predicted occurrence (blue). Collection locations are represented by dots; dots may represent more than one sample collected at those coordinates. White dots represent locations where at least one AIV-positive sample was collected; the black dots are locations where only AIV-negative samples were collected. The predictive model was constructed by applying the Random Forests algorithm to 21,149 wild bird data points and 51 predictor variables.

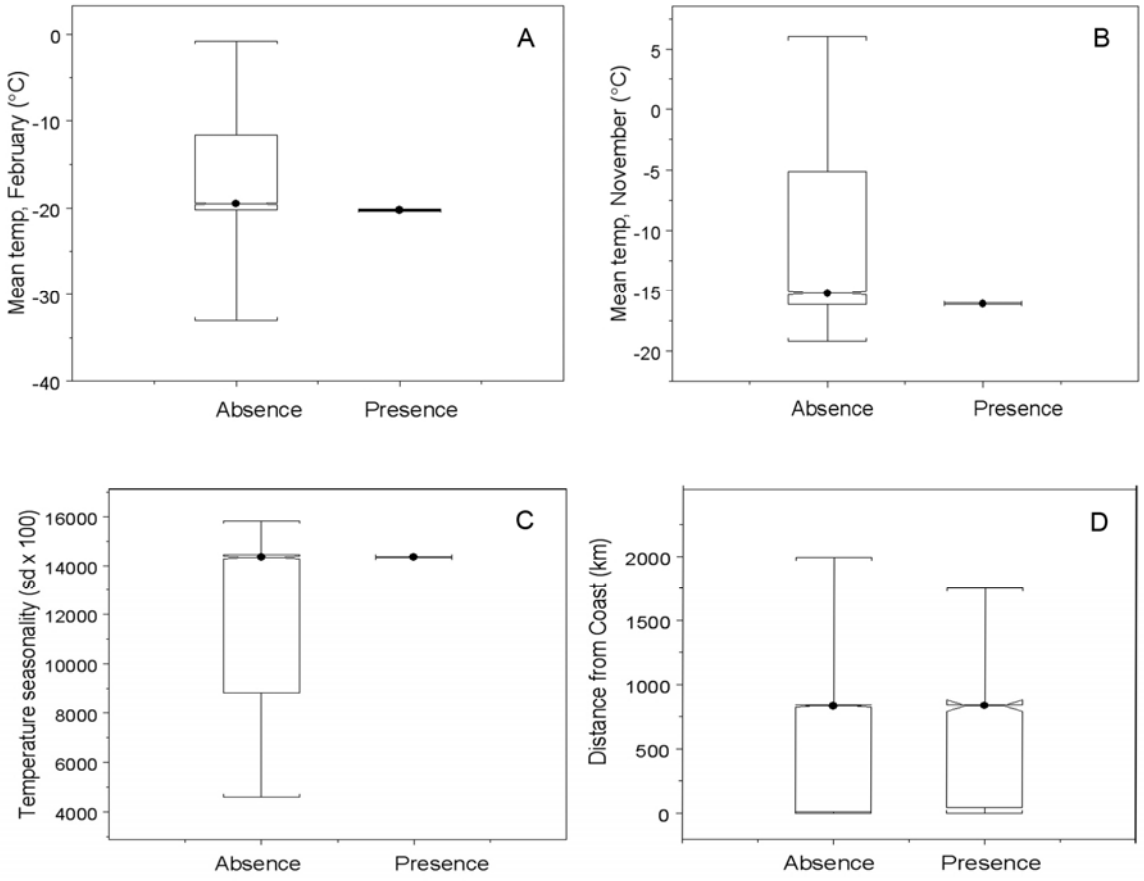

Figure 2.2. Notched box plots for important variables. Important variable values are grouped by samples which tested negative for avian influenza virus (AIV; Absence) and AIV-positive samples (Presence). Mean temperature in February (A), Mean temperature in November (B), Temperature seasonality (C), and distance from coast (D) were the highest contributors to model accuracy as calculated by Random Forests. Box plots display the median values (central dot), 95% confidence intervals (notch around the median), and the maximum and minimum range values (whiskers).

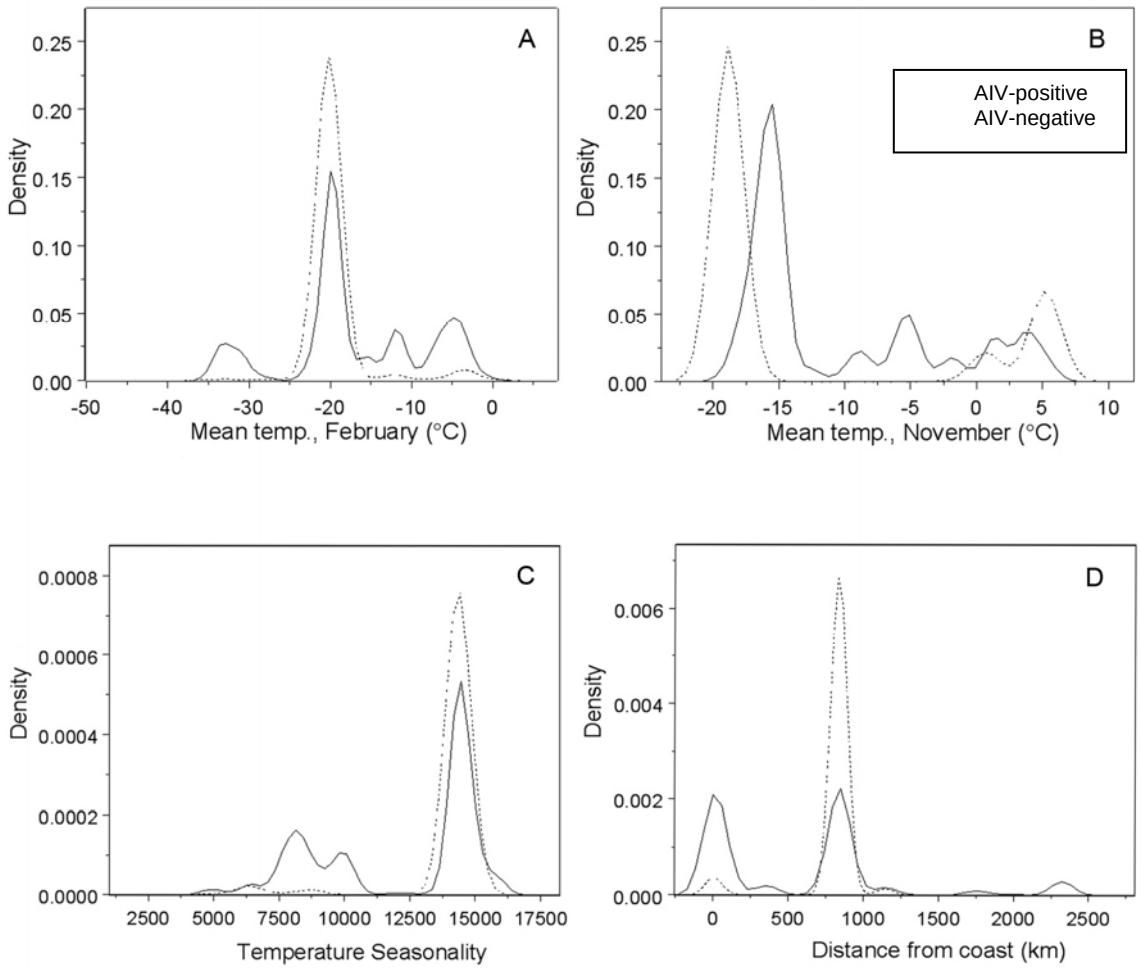

Figure 2.3. Histogram density plots for important variables. Mean temperature in November (B), Temperature seasonality (C), and distance from coast (D) were the highest contributors to model accuracy as calculated by Random Forests. Density lines for cases that tested positive for avian influenza virus (AIV) are represented by dashed lines, AIV-negative by solid lines. Peaks in AIV-positive cases that exceed AIV-negative cases indicate a range of values for this predictor variable that are correlated with AIV-positivity.

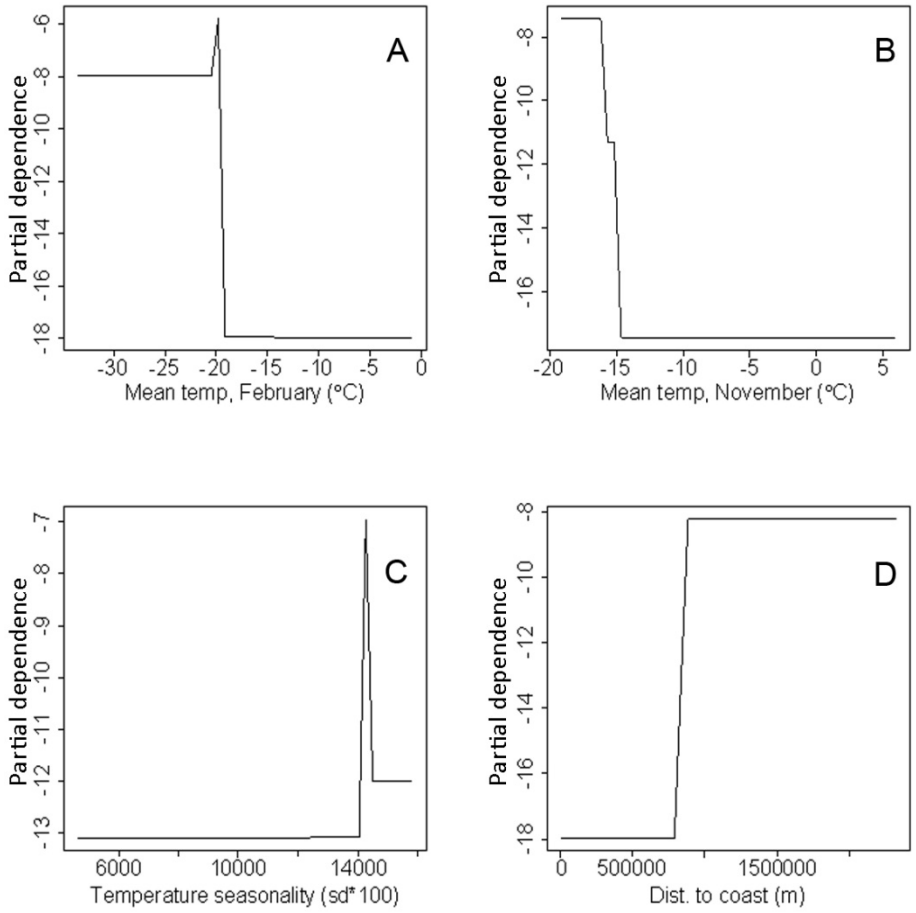

Figure 2.4. Partial dependence plots for important variables. Mean temperature in November (B), Temperature seasonality (C), and distance from coast (D) were the highest contributors to model accuracy as calculated by Random Forests. Plots show the partial dependence of avian influenza virus (AIV)-positivity on each predictor variable over the averaged effects of all other predictors. Partial dependence is best interpreted as an index of the importance of values within a variable's range and is best understood by examining general patterns in relation to the values of the predictor variable rather than the specific values of partial dependence. For example, in Figure 4A, a mean temperature in February below -20 °C is most correlated with AIV-positivity; ranges above -20 °C show very low correlation.

LITERATURE CITED

1. The World Bank - East Asia and Pacific Region (2005) *Spread of avian flu could affect next year's economic outlook*. Excerpted from Nov. 2005 East Asia Update - Countering Global Shocks.

2. Lignon, B.L.: Avian influenza virus H5N1: a review of its history and information regarding its potential to cause the next pandemic. *Seminars in Pediatric Infectious Diseases*, 16(3): 26-35 (2005).

3. World Health Organization. *Cumulative number of confirmed human cases for avian influenza A(H5N1) reported to WHO, 2003-2011* (2011). Available from: http://www.who.int/influenza/human_animal_interface/EN_GIP_20111010CumulativeNumberH5N1cases.pdf.

4. Munster, V.J., Baas, C. Lexmond, P., Waldenström, J., *et al*. : Spatial, temporal, and species variation in prevalence of influenza A viruses in wild migratory birds. *PLoS Pathogens*, 3(5): 630-38 (2007).

5. Krauss, S., Walker, D., Pryor, S.P., et al.: Influenza A viruses of migrating wild aquatic birds in North America. *Vector Borne and Zoonotic Diseases*, 4(3): 177-89 (2004).

6. Webster, R., Bean, W.J., Gorman, W.T., et al.: Evolution and ecology of influenza A viruses. *Microbiological Reviews*, 56(1): 152-79 (1992).

7. World Organization for Animal Health (OIE), *Avian Influenza* in *Manual of Diagnostic Tests and Vaccines for Terrestrial Animals (version adopted May 2012).* OIE, Paris (2012).

8. Hanson, B.A., Stallknecht, D.E., Swayne, D.E., et al.: Avian influenza viruses in Minnesota ducks during 1998-2000. *Avian Diseases*, 47(s3): 867-71 (2003).

9. Kaleta, E., Hergarten, G., Yilmaz, A.: Avian influenza A viruses in birds of the order Psittaciformes: reports on virus isolations, transmission experiments and vaccinations and initial studies on innocuity and efficacy of oseltamivir in ovo. *Deutsche tierarztliche Wochenschrift*, 114(7): 260-67 (2007).

10. Goyal, S. M., Jindal, N., Chander, Y., et al.: Isolation of mixed subtypes of influenza A virus from a bald eagle (*Haliaeetus leucocephalus*). *Virology Journal*, 7(1): 174 (2010).

11. Stallknecht, D. and S. Shane: Host range of avian influenza in free-living birds. *Veterinary Research Communications*, 12(2-3): 125-141 (1988).

12. Abolnik, C., Gerdes, G., Sinclair, M., et al.: Phylogenetic analysis of influenza A viruses (H6N8, H1N8, H4N2, H9N2, H10N7) isolated from wild birds, ducks, and ostriches in South Africa from 2007 to 2009. *Avian Diseases*, 54(S1): 313-22 (2010).

13. Laudert, E.A., V. Sivanandan, and D.A. Halvorson: Effect of intravenous inoculation of avian infleunza virus on reproduction and growth in mallard ducks. *Journal of Wildlife Diseases*, 29(4): 523-26 (1993).

14. Slemons, R.D. and B.C. Easterday: Virus replication in the digestive tract of ducks exposed by aerosol to type-A influenza. *Avian Diseases*, 22(3): 367-77 (1978).

15. Kida, H., R. Yanagawa, and Y. Matuoka: Duck influenza lacking evidence of disease signs and immune response. *Infection and Immunity* 30(2): 547-53 (1980).

16. Pasick, J., Berhane, Y., Hisanaga, T., et al.: Diagnostic test results and pathology associated with the 2007 Canadian H7N3 highly pathogenic avian influenza outbreak. *Avian Diseases*, 54(s1): 213-19 (2010).

17. Davison, S., R.J. Eckroade, and A.F. Ziegler: A Review of the 1996–98 Nonpathogenic H7N2 avian influenza outbreak in Pennsylvania. *Avian Diseases*, 47(s3): 823-827 (2003).

18. Nestorowicz, A., Kawaoka, Y., Bean, W., et al.: Molecular analysis of the hemagglutinin genes of Australian H7N7 influenza viruses: role of passerine birds in maintenance or transmission? *Virology*, 160(2): 411-18 (1987).

19. Sinclair, P.N.T.M. and B. Ganzevoort: Risk factors for seropositivity to H5 avian influenza virus in ostrich farms in the Western Cape Province, South Africa. *Preventive Veterinary Medicine*, 86(1-2): 139-52 (2008).

20. Halvorson, D.A., C.J. Kelleher, and D.A. Senne: Epizootiology of avian influenza: effect of season on incidence in sentinel ducks and domestic turkeys in Minnesota. *Applied and Environmental Microbiology*, 49(4): 914-19 (1985).

21. CEIRS. *"Centers of Excellence for Influenza Research and Surveillance (CEIRS)". National Institute of Allergy and Infectious Diseases.* Available from: http://www3.niaid.nih.gov/LabsAndResources/resources/ceirs/introduction.htm.

22. Derksen, D. *Progress report: assessment of virus movement across continents: using Northern Pintails (Anas acuta) as a test* (2008). Available from: http://alaska.usgs.gov/science/biology/avian_influenza/pdfs/Assessment_of_virus_movment_progress_report_2008.pdf.

23. Ramey, A. M., Pearce, J. M., Ely, C. R., et al.: Transmission and reassortment of avian influenza viruses at the Asian–North American interface *Virology*, 406(2): 352-59 (2010).

24. Tran, A., Goutard, F., Chamaille, L., et al.: Remote sensing and avian influenza: A review of image processing methods for extracting key variables affecting avian influenza virus survival in water from Earth Observation satellites. *International Journal of Applied Earth Observation and Geoinformation*, 12(1): 1-8 (2010).

25. Peterson, A. T., Sanchez-Cordero, V., Beard, C. B., et al.: Ecological niche modeling and potential reservoirs for Chagas disease, Mexico. *Emerging Infectious Diseases*, 8(7): 662-67 (2002).

26. Mak, S., M. Morshed, and B. Henry: Ecological niche modeling of Lyme Disease in British Columbia, Canada. *Journal of Medical Entomology*, 47(1): 99-105 (2010).

27. Moffett, A., N. Shackelford, and S. Sarkar: Malaria in Africa: vector species' niche models and relative risk maps. *PLOS ONE*, 2(9): 1-14 (2007).

28. Peterson, A.T., D.A. Vieglais, and J.K. Andreasen: Migratory birds modeled as critical transport agents for West Nile Virus in North America. *Vector Borne and Zoonotic Diseases*, 3(1): 27-37 (2003).

29. Runstadler, J. A., Happ, G. M., Slemons, R. D., et al.: Using RRT-PCR analysis and virus isolation to determine the prevalence of avian influenza virus infections in ducks at Minto Flats State Game Refuge, Alaska, during August 2005. *Archives of Virology*, 152(10): 1901-1910 (2008).

30. Spackman, E., Senne, D., Myers, T., et al.: Development of a real-time reverse transcriptase PCR assay for type A influenza virus and the avian H5 and H7 hemagglutinin subtypes. *Journal of Clinical Microbiology*, 40(9): 3256-60 (2002).

31. Hijmans, R. J., Cameron, S. E., Parra, J. L., et al.: Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25(15): 1965-78 (2005).

32. Lehner, B. and P. Döll: Development and validation of a global database of lakes, reservoirs and wetlands. *Journal of Hydrology*, 296(1-4): 1-22 (2004).

33. Sanderson, E. W., Jaiteh, M., Levy, M. A., et al.: The Human Footprint and the Last of the Wild. *BioScience*, 52(10): 891-904 (2002).

34. Center for International Earth Science Information Network (CIESIN) &Columbia University; and Centro Internacional de Agricultura Tropical (CIAT). *Gridded Population of the World Version 3 (GPWv3): Population Density Grids,* (2005).

35. Robinson, T.P., G. Franceschini, and G. Wint: The Food and Agriculture Organization's Gridded Livestock of the World. *Veterinaria Italiana*, 43(3): 745-51 (2007).

36. Cutler, D. R., Edwards, T. C., Jr., Beard, K. H., et al.: Random Forests for classification in ecology. *Ecology*, 88(11): 2783-2792 (2007).

37. Breiman, L.: Random Forests. *Machine Learning*, 45(1): 5-32 (2001).

38. Pontius Jr., R.G. and L.C. Schneider: Land-cover change model validation by an ROC method for the Ipswich watershed, Massachusetts, USA. *Agriculture, Ecosystems and Environments*, 85(239-248 (2001).

39. Liaw, A. and M. Wiener: Classification and regression by RandomForest. *R News*, 2(3): 18-22 (2002).

40. R Development Core Team, *R: A language and environment for statistical computing*. 2010, R Foundation for Statistical Computing: Vienna.

41. Elith, J., C. Graham, and NCEAS working group: Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29(1): 29-51 (2006).

42. Craig, D. and F. Huettmann, *Chapter IV, Using "blackbox" algorithms such as TreeNet and RandomForest for data-mining and for finding meaningful patterns, relationships and outliers in complex ecological data* in *Intelligent Data Analysis: Developing New Methodologies through Pattern Discovery and Recovery.* IGI Global, Hershey (2008).

43. Peterson, A.T. and R.A.J. Williams: Risk mapping of highly pathogenic avian influenza distribution and spread. *Ecology and Society*, 13(2): 15 (2008).

44. Gilbert, M., Xiao, X., Pfeiffer, D. U., et al.: Mapping H5N1 highly pathogenic avian influenza risk in Southeast Asia. *Proceedings of the National Academy of Sciences of the United States of America*, 105(12): 4769-74 (2008).

45. Moffett, A., N. Shackelford, and S. Sarkar: Malaria in Africa: vector species' niche models and relative risk maps. *PLoS ONE*, 2(9): e824 (2007).

46. Peterson, A. T., Lash, R. R., Carroll, D. S., et al.: Geographic potential for outbreaks of Marburg hemorrhagic fever. *The American Journal of Tropical Medicine and Hygiene*, 75(1): 9-15 (2006).

47. Adhikari, D., A. Chettri, and S.K. Barik: Modelling the ecology and distribution of highly pathogenic avian influenza (H5N1) in the Indian subcontinent. *Current Science*, 97(1): 72-78 (2009).

48. Chace, J.F. and J.J. Walsh: Urban effects on native avifauna: a review. *Landscape and Urban Planning*, 74(1): 46-69 (2006).

49. Center for International Earth Science Information Network (CIESIN), Columbia University, and Centro Internacional de Agricultura Tropical (CIAT). *Gridded Population of the World Version 3 (GPWv3): Population Density Grids* (2005). Available from: http://sedac.ciesin.columbia.edu/gpw.

CHAPTER 3
A global model of avian influenza prediction in wild birds: the importance of northern regions[2]

ABSTRACT

Avian influenza virus (AIV) is enzootic to wild birds, which are its natural reservoir. The virus exhibits a large degree of genetic diversity and most of the isolated strains are of low pathogenicity to poultry. Although AIV is nearly ubiquitous in wild bird populations, highly pathogenic H5N1 subtypes in poultry have been the focus of most modeling efforts. To better understand viral ecology of AIV, a predictive model should 1) include wild birds, 2) include all isolated subtypes, and 3) cover the host's natural range, unbounded by artificial country borders. As of this writing, there are few large-scale predictive models of AIV in wild birds. We used the Random Forests algorithm, an ensemble data-mining machine-learning method, to develop a global-scale predictive map of AIV, identify important predictors, and describe the environmental niche of AIV in wild bird populations. The model has an accuracy of 0.79 and identified northern areas as having the highest relative predicted risk of AIV-positivity. Important predictor variables included high annual precipitation, low mean temperature in June (<15 ^{0}C), and low mean temperature in April (<0 ^{0}C). This study is the first global-scale model of low-pathogenicity avian influenza in wild birds and underscores the importance of largely unstudied northern regions in the persistence of AIV.

---

[2] Herrick KA, Huettmann F, Lindgren MA: A global model of avian influenza prediction in wild birds: the importance of northern regions. *Vet Res*, in review.

INTRODUCTION

The influenza viruses that caused the four deadliest human pandemics of the past century (1918, 1957, 1968, 2009) contained gene segments from avian influenza acquired through recent reassortment events (reviewed in [1]). Influenza is thought to have originated in wild birds, and waterfowl are considered the primary reservoir. Avian influenza virus (AIV) most commonly infects *Anseriformes*, *Passeriformes*, and *Charadriiformes* in wild populations, particularly family *Anatidae* [2]. The Asian strains of the highly pathogenic H5N1 AIV subtype in poultry have received the most attention because of economic losses caused by this subtype, the virus's transmissibility from chicken to human [3, 4], and fears over a new human influenza pandemic [5]. However, HPAI H5N1 is not the only strain with pandemic potential: cases of human infection with interspecies H7 and H9 subtypes have been reported [6, 7] and others, such as H6, can be highly pathogenic to poultry. The vast majority of influenza strains are of low pathogenicity to poultry, but because AIV is a virus of great diversity with the potential for rapid evolution [8], the full range of its variation should be considered rather than just focusing on a single strain or subtype. A massive reservoir of genetic diversity for potential reassortment of AIV exists in wild bird populations, from which nearly all combinations of hemagglutinin (H) and neuraminidase (N) subtypes have been isolated [9-11].

A number of ongoing surveillance projects record the subtypes of AIV isolated from wild birds [9, 10, 12, 13]. Large cooperative databases, such as the Influenza Research Database (IRD), curate the surveillance efforts of multiple institutions. IRD

provides an opportunity to apply predictive modeling to AIV on a global scale. In the prediction and risk assessment of infectious diseases, geographic information systems (GIS) and predictive modeling techniques are important tools [14]. Predictive models of Chagas disease [15], malaria [16], leishmaniasis [17], and Lyme disease [18] have been used to map disease prevalence and identify important factors contributing to risk. Several models have assessed risk factors for H5N1 in domestic poultry and produced high resolution models for India [19], Vietnam [20], and China [21]. One predictive map of multiple species of wild birds and subtypes of AIV was developed for the continental United States [22], and one for flyways adjacent to the Pacific Rim [23].

As of this writing, there are no global-scale predictions of LPAI in wild birds. The development of a global model could be an important tool in the management and risk assessment of AIV in the interest of public and animal health. Our model extends the value of AIV surveillance efforts by using the data for predictive purposes, not simply for descriptive purposes. In addition, a global model encompasses the distribution range of important reservoir species, many of which travel vast distances [24] in cross-continental migratory journeys to and from breeding grounds each year. We used ensemble data-mining machine-learning methods to 1) identify important predictor variables, 2) quantitatively describe the environmental niche of AIV in wild bird populations, and 3) develop a near-global-scale predictive map (excluding Antarctica) of AIV based on this described niche.

MATERIALS AND METHODS

Wild bird data

Sample data points of AIV-negative and AIV-positive data for wild birds were obtained from the Influenza Research Database online at http://www.fludb.org [25]. This dataset spans five years (2005-2010) of surveillance data providing georeferenced collection coordinates for each sample, species name, AIV-positive or –negative status (determined by the collecting institution), viral subtype (where available), and many other collection specifics. We did not distinguish between high- and low-pathogenicity AIV strains in our dataset. We groomed the database to remove samples from domestic species and samples from unidentified species (listed as "Unknown"). In addition, this version of the database contained many instances where the latitude and longitude values were inverted; we examined each point for a match between GPS coordinates and collection location, and corrected it if the error was obvious or removed it if uncertainty remained. The remaining data were made up of 59,978 georeferenced points with an AIV-positive prevalence of 3.3%. We randomly divided the data points into two pools for training the model (80%) and testing the model (20%) using MS Excel and imported both sets as point layers into ArcMap v.10.0 (Esri). A table of all bird species represented in the original database, including the total number of samples collected and the number that tested positive for AIV are presented as Appendix F.

Environmental variable layers

Forty-two predictor variable layers for ArcGIS were acquired from open source projects and included bioclimatic, geographic, and anthropogenic variables (Table 3.1). The

extent of this model is bounded by these data layers, which exclude Antarctica. Bioclimatic variables included mean temperature for each month, for quarters (e.g. wettest quarter), and annual means for precipitation and temperature. A number of time-dependent variables were included (i.e. mean monthly temperatures in January through December) and were manipulated in order to maintain their relevance to collection locations in the Southern Hemisphere. For points with negative latitude values, time-dependent variables were shifted by 6 months, such that months were correctly associated with the austral seasons.

Geographic variables included elevation, which has been identified as an important factor in other AIV models [19], and lakes, rivers, and wetlands, which are important to waterfowl. We calculated some layers from existing predictor variables using the Spatial Analyst Tool in ArcMap. The distances from fresh water features and coastline were calculated using the Euclidean Distance Tool. Slope was calculated from elevation and aspect was in turn calculated from slope. Anthropogenic variables included the Human Influence and Human Footprint indices, which reflect the extent of human manipulation, infrastructure, and population density [26]. Due to the importance of chickens and pigs in the transmission of AIV to humans, we included predicted poultry and pig densities [27, 28]. Not all layers included Antarctica, so the entire continent was excluded from the study area (layers trimmed at -57^0 latitude) to prevent biases in calculation. We then used the Geospatial Modeling Environment (GME; [29]) to intersect, or extract the values of the predictor variables at the same geographic coordinates as the sample data points. GME adds the values of each predictor variable to

the database as an additional column. The intersected database is then imported into ArcMap for visualization. Layers and metadata are stored at and can be obtained from the Ecological Wildlife Habitat Analysis of the Land- and Seascape (EWHALE) Lab at the University of Alaska Fairbanks (UAF).

Defining the outbreak niche

We used the Random Forests algorithm [30], an ensemble data-mining machine-learning method, to identify the variables that best predicted the AIV-positive niche. We chose this particular algorithm because it is a powerful method of data-mining that performs with equal or superior accuracy to other algorithms (such as TreeNet, MARS, and Regression Tree Analysis) when used in ecological prediction [31, 32]. Random Forests is relatively immune to overfitting and noise [30], which is a valuable feature when many similar predictor variables are incorporated. In addition, Random Forests ranks predictor variables by their contribution to model accuracy and the Variable Importance Scores (VIS) are normalized to the highest scoring variable. Using the pool of training data, we ran the Random Forests analysis method for classification trees in Salford Predictive Miner (Salford Systems) with the following settings: class weights were balanced to up-weight the smaller number of AIV-positive samples against AIV-negative samples; the number of trees was set to 500; and seven predictors were used at each node [30].

The top five variables with the highest VIS were chosen for further examination. To compare the number of AIV-positive and -negative samples taken across each variable's range of values we plotted density using Spotfire S+ (TIBCO, v.8.2). The ranges within which peaks occur suggest underlying mechanisms, which may be driving

AIV outbreaks. Partial dependence plots were produced using the "partialPlot()" command in the randomForest package [33] in R statistical programming language [34]. Partial dependence can be thought of as an index summarizing the quantified relationship of a predictor with the response variable after averaging the noise of non-relevant predictors [35]. Partial dependence plots can be useful in illustrating general trends in model accuracy's dependence on predictors. The partial dependence of a variable's effect is best understood by examining general patterns in relation to the values of the predictor variable rather than the specific values of partial dependence.

Predictive map

To predict the relative occurrence of AIV in unsampled areas, we applied the model to a grid of points spaced 100 km apart and calculated a predicted value for each point. Random Forests expresses the predicted occurrence of AIV as a Relative Occurrence Index (ROI) rather than a probability score [36]. In ArcMap, we applied the Inverse Distance Weighted Tool (IDW) to interpolate these ROI values between the points, and generated a map of predicted AIV outbreak locations.

To evaluate the performance of the model, we used the Receiver Operating Characteristic (ROC) curve by plotting true positive points (AIV-positive status) against false positives and calculating the resulting area under the curve (AUC) with the program ROC_AUC [37]. An AUC value of 0.5 means a model accuracy of 50% in predicting positives and is no better than the random assignment of positive or negative status. An AUC value of 1.0 shows the model accurately classified 100% of points. If AUC exceeds the critical value of 0.7, the model has high predictive power [38]. To evaluate accuracy,

the model was applied to the pool of testing points. A ROC curve was calculated for these points using their predicted ROI value against their experimental AIV-positive status.

RESULTS

Important predictor variables

Annual precipitation, mean temperature in June, and mean temperature in April were the most important predictor variables with VIS of 100, 85.2, and 76.1, respectively. Predictor variables with VIS above 50 were split almost equally between precipitation measurements and the mean temperatures in November (austral May), the driest quarter (3 month period), and annual mean temperature (Table 3.1). In the density plots (Fig. 3.1A-E) the relative frequency of sampling was approximated by the density of the AIV-negative group of samples (represented by the solid black line); the range of values over which sampling occurred was inferred from the AIV-negative group. In general, the lack of perfect correspondence between AIV-positive (dotted red line) and AI–negative groups showed that there were unequal densities of AIV-positivity across the sampling range. Thus AIV-positive samples did not occur at the same relative frequency as sampling effort. The ranges where the density of AIV-positive samples exceeded those of AIV-negative samples imply conditions correlated with AIV-positivity. In the case of annual precipitation (Fig. 3.1A), moderate (1400 mm) and very low (~0 mm) values were correlated with AIV-positivity. The partial dependence of AIV-positivity on annual precipitation exhibited a similar trend: very high dependence from 0 mm to 500 mm, a

trough, and then moderately high dependence at values over 1000 mm (Fig. 3.2A). These patterns imply that areas of low annual precipitation are most correlated with AIV-positivity, although areas of relatively high annual precipitation show some correlation as well. Areas of very low and high mean temperatures in June and April (austral December and October)were correlated with AIV-positivity (Figs. 1B, C), while areas of moderate temperature were not. June and April displayed similar patterns with a strong peak at the high range of sampling (~28 ^{0}C and 30 ^{0}C, respectively) and at the lowest ranges (~10 ^{0}C and 0 ^{0}C, respectively). Examination of partial dependence revealed that AIV-positivity was high at the lowest temperatures, dropped sharply at moderate temperatures, and gradually increased at the higher end of the range (Figs. 2B, C). Thus areas with low temperatures in June and April were correlated with AIV-positive samples. Precipitation of the driest quarter displayed one peak at 50 mm where AIV-positives had a higher density than AIV-negatives (Fig. 3.1D). However, while the partial dependence was high at this value, it appears as a lone spike in an area of low partial dependence. Partial dependence on precipitation increases above 150 mm and reaches high levels above 250 mm (Fig. 3.2D). While the highest density of AIV-positives occurred at relatively low annual precipitation, partial dependence was highest at the highest range during the driest quarter, which may reflect a low seasonality or variation in rainfall during the year. Mean temperature in November was correlated with AIV-positivity at low and high values (Figs. 1E, 2E). Highest partial dependence occurred at the lowest ranges (< -20 ^{0}C). Based on the important predictor variables, the niche of AIV-positive samples in this study was described as regions of low annual rainfall and low temperatures. There

appears to be a secondary niche that described regions of high precipitation and higher temperatures.

Ecological niche model

Random Forests produced a robust ecological niche model for AIV in wild birds and identified important predictor variables. The model had an ROC/AUC of 0.79 on the training points and 0.76 on the testing points, lending high confidence to its prediction of the relative occurrence of AIV in wild birds on a global scale. Northern areas had the highest values of ROC and temperate regions had the lowest (Fig. 3.3). Interestingly, an equatorial band of relatively high predicted occurrence was observed, which may reflect regions characterized by the secondary niche.

DISCUSSION

While much of AIV modeling has focused on low-latitude regions and HPAI H5N1, we demonstrated that northern regions are important when all strains of AIV and wild reservoir species are taken into account. By creating a global-scale model, we identified important areas of high predicted occurrence that were missed by AIV models for temperate and sub-tropical regions. Small, local models are vital developing strategies for managing acute outbreaks of specific diseases. However, a global scale perspective is necessary for AIV because, unlike other diseases, is carried by a host that is capable of migrating long distances and potentially infecting others along its path. Furthermore, a model that excludes wild birds, which are the natural reservoir for the virus, neglects the source of gene segments for future infections and potential pandemic strains.

Our model represents the first global-scale predictive map of AIV in wild birds. Using available global AIV data, we identified northern areas as having the highest relative predicted risk of outbreak. Important predictor variables included low temperatures and low annual precipitation. Cold winters and low rainfall may represent continental climates at high latitudes. Areas with these types of climatic conditions include landscapes in Siberia, the Russian Far East, Mongolia, and northern Canada, all of which had high indices of relative occurrence of AIV. Similar conditions at lower latitudes may be created by high elevation, such as the climate of the Tibetan Plateau, which also had a high score. The partial dependence of AIV-positivity on rainfall was bimodal and peaked at very low and high values. This apparently contradictory finding that extremes in rainfall were correlated with AIV-positivity may be explained through laboratory studies of transmission and persistence of the virus. Aerial, non-contact transmission of influenza between guinea pigs was most efficient below 35% relative humidity [39]. At low relative humidity and temperature (~6 ºC, < 46 % RH), virus persisted over two weeks on metal, glass, and in soil [40]. Thus, we expect dry climate to be conducive to the aerial spread of virus. Wet conditions and low temperatures were also conducive to viral persistence: the virus remains viable nearly ten times longer in 17 ºC water than 28 ºC water [41]. At low temperatures and high relative humidity (~7 ºC, ~88 % rh), the virus persisted over two weeks in chicken feces [40]. Low temperature is the common factor in these studies. While low relative humidity contributes to transmission and persistence on smooth surfaces, the virus also remains viable in water and damp materials such as bird feces. As the virus is transmitted efficiently in water, either through

the fecal-oral route [41] or via tracheal shedding [42], dabbling ducks (such as *Anatidae*) in cool northern regions may be at increased risk of contracting AIV from the environment.

Our findings differed from other AIV models in the importance and range of anthropogenic variables. In our model, anthropogenic factors were represented by human population density as well as the Human Influence Index and the Human Footprint Index [43], which are indices calculated based on human population density, land transformation, transportation infrastructure, and electrical power infrastructure. All the anthropogenic variables received very low VIS with human population density scoring the highest at 29.3. Previous models identified high human population density and high farming intensity (especially rice cropping and aquaculture) as important predictors [19, 20, 44]. The niche they described is characterized as having a high human population, high level of anthropogenic disturbance, and the high annual temperature and humidity of the sub-tropical climates for which the models were designed (i.e. Bangladesh, Vietnam, and Thailand). However, these studies were specific to HPAI H5N1 in poultry. While the one North American model in wild birds identified low minimum temperatures, with which our model was consistent, they also identified the amount of cropland as an important factor [22]. In general, our model did not predict high occurrence of AIV in the continental United States when compared to northern regions, which have not been modeled previously.

Our model demonstrated a novel use of surveillance data that goes beyond the yearly reporting of infected species and viral subtypes isolated. The application of

environmental data, GIS, and machine-learning extends the usefulness of surveillance results. However, the prediction of relative occurrence presented here is not a final, definitive map of avian influenza in wild birds, but rather an initial attempt that demonstrates that a useful signal can be gleaned from the noise found in a global dataset. Indeed, it serves to highlight shortcomings in available data. In particular, nearly all data were collected in the Northern Hemisphere. In addition, this Northern Hemispheric niche could then be tested on southward-migrating birds to see if the same predictions are applicable. A predominance of *Anatidae* could create a spatial bias for northern regions and a temporal bias for summer months if most sampling is carried out during summer breeding season at high latitudes. However, if one uses the mean temperature in November as a proxy for latitude, there appears instead to be a strong temperate bias in collection with AIV-positive peaks occurring to either side. The bifurcate niche evident here is an interesting topic for future analysis. The mechanisms responsible for this niche require further investigation in order to clarify how the important bioclimatic variables contribute to AIV-positivity.

While ongoing surveillance is important to understanding the dynamics of AIV, efforts should include wilderness areas, such as Siberia, that have received less attention. Models such as this one could receive additional fine-tuning if these results were to guide future sampling efforts in regions of high predicted occurrence, much of which remains unsampled. As both AIV-positive and AIV-negative data are incorporated into this model, all results from prediction-guided sampling strengthen the prediction, even if only a small percentage of AIV-positive samples are isolated. Given the sheer quantity of data

collected by long term surveillance efforts, an unprecedented opportunity exists to produce future models of greater accuracy. If data were curated and publically available, models could be treated as transparent, replicable science experiments. Improved global scale models could not only increase the understanding of viral ecology, but also serve to guide the management of influenza risk policy for the benefit of public health on a global scale. A global model of AIV must be a collaborative effort and we hope this initial attempt encourages greater cooperation and data-sharing among members of the AIV research community.

ACKNOWLEDGEMENTS

We thank IRD for providing the wild bird data used in this project. The Influenza Research Database Bioinformatics Resource Center has been wholly funded with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN266200400041C. We thank the open source projects that provided the predictor variables used in this project. We thank Dr. Abby Powell and BIOL 644 for helpful comments on the manuscript. KAH was partially funded by the Biology & Wildlife Department (UAF). The funding body played no role in data collection or analysis; in the writing of the manuscript; or its decision to submit for application. ML donated his time. This is EWHALE Lab Publication #111.

TABLES

Table 3.1. The predictor variables used by the Random Forests algorithm to create a global prediction map for avian influenza virus in wild birds. Variables are listed in order of their Variable Importance Score (VIS) or relative contribution to model accuracy as calculated by Random Forests.

Predictor Variable	Details	VIS	Project Source
Annual precipitation (mm)	In mm; 30 arc-seconds, 1 km spatial resolution	100.0	WorldClim [45]
Mean temperature, June (°C)	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	85.2	WorldClim
Mean temperature, April (°C)	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	76.1	WorldClim
Precipitation of driest quarter (mm)	In mm; 30 arc-seconds, 1 km spatial resolution	68.4	WorldClim
Mean temperature, November (°C)	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	64.9	WorldClim
Precipitation seasonality	In mm; 30 arc-seconds, 1 km spatial resolution	63.1	WorldClim
Mean temperature of driest quarter (°C)	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	62.5	WorldClim
Annual mean temperature (°C)	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	54.8	WorldClim
Mean temperature, February (°C)	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	49.7	WorldClim
Mean temperature, January (°C)	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	45.7	WorldClim
Temperature seasonality	Standard deviation x 100	45.0	WorldClim
Precipitation of wettest quarter (mm)	In mm; 30 arc-seconds, 1 km spatial resolution	42.9	WorldClim
Mean temperature, December (°C)	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	38.6	WorldClim
Maximum temperature of warmest month (°C)	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	38.0	WorldClim
Precipitation of driest month (mm)	In mm; 30 arc-seconds, 1 km spatial resolution	37.5	WorldClim
Mean temperature, October (°C)	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	36.1	WorldClim
Mean temperature, September (°C)	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	32.6	WorldClim
Precipitation of coldest quarter (mm)	In mm; 30 arc-seconds, 1 km spatial resolution	32.3	WorldClim

Table 3.1 continued

Predictor Variable	Details	VIS	
Population density (persons/km^2)	Population density for 2010, 2.5' resolution, persons/km^2	29.3	Gridded Popn of the World, v.3 [46]
Mean temperature of coldest quarter (°C)	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	28.4	WorldClim
Mean temperature, July (°C)	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	28.3	WorldClim
Isothermality (°C)	(Mean Diurnal Range/Temperature Annual Range); In ^{o}C; 30 arc-seconds, 1 km spatial resolution	26.6	WorldClim
Mean temperature of wettest quarter (°C)	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	25.2	WorldClim
Mean diurnal range (°C)	In ^{o}C, (mean of monthly temperature(max – min))	24.3	WorldClim
Mean temperature, August (°C)	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	22.7	WorldClim
Mean temperature, March (°C)	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	21.0	WorldClim
Temperature annual range (°C)	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	18.4	WorldClim
Elevation (m)	In m; 30 arc-seconds, 1 km spatial resolution	18.3	WorldClim
Predicted pig density (per km^2)	Animal density/km^2 ; 3 min of arc	18.0	Gridded livestock of the world [47]
Mean temperature of warmest quarter (°C)	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	17.0	WorldClim
Precipitation of wettest month (mm)	In mm; 30 arc-seconds, 1 km spatial resolution	15.0	WorldClim
Distance from coast (m)	Calculated from coastline	12.6	Esri
Precipitation of warmest quarter (mm)	In mm; 30 arc-seconds, 1 km spatial resolution	12.1	WorldClim
Minimum temperature of coldest month (°C)	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	11.2	WorldClim
Human footprint index	Percentage of relative human influence (0-100)	10.4	Last of the Wild [26]
Distance from rivers, lakes, or wetlands	Calculated from combined large and small lake polygons, and lakes and wetlands grid	9.7	Global Lakes and Wetlands Database [48]
Slope	Calculated from elevation	9.0	Esri
Human influence index	Summative index of human disturbance (0-72)	8.8	Last of the Wild
Predicted poultry density (per km2)	Animal density/km^2 ; 3 min of arc	7.9	Gridded livestock of the world [47]

Table 3.1 continued

Predictor Variable	Details	VIS	
Aspect	positive degrees from 0 to 359.9, measured clockwise from north; calculated from slope	6.7	Esri
Mean temperature, May (°C)	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	5.4	WorldClim

FIGURES

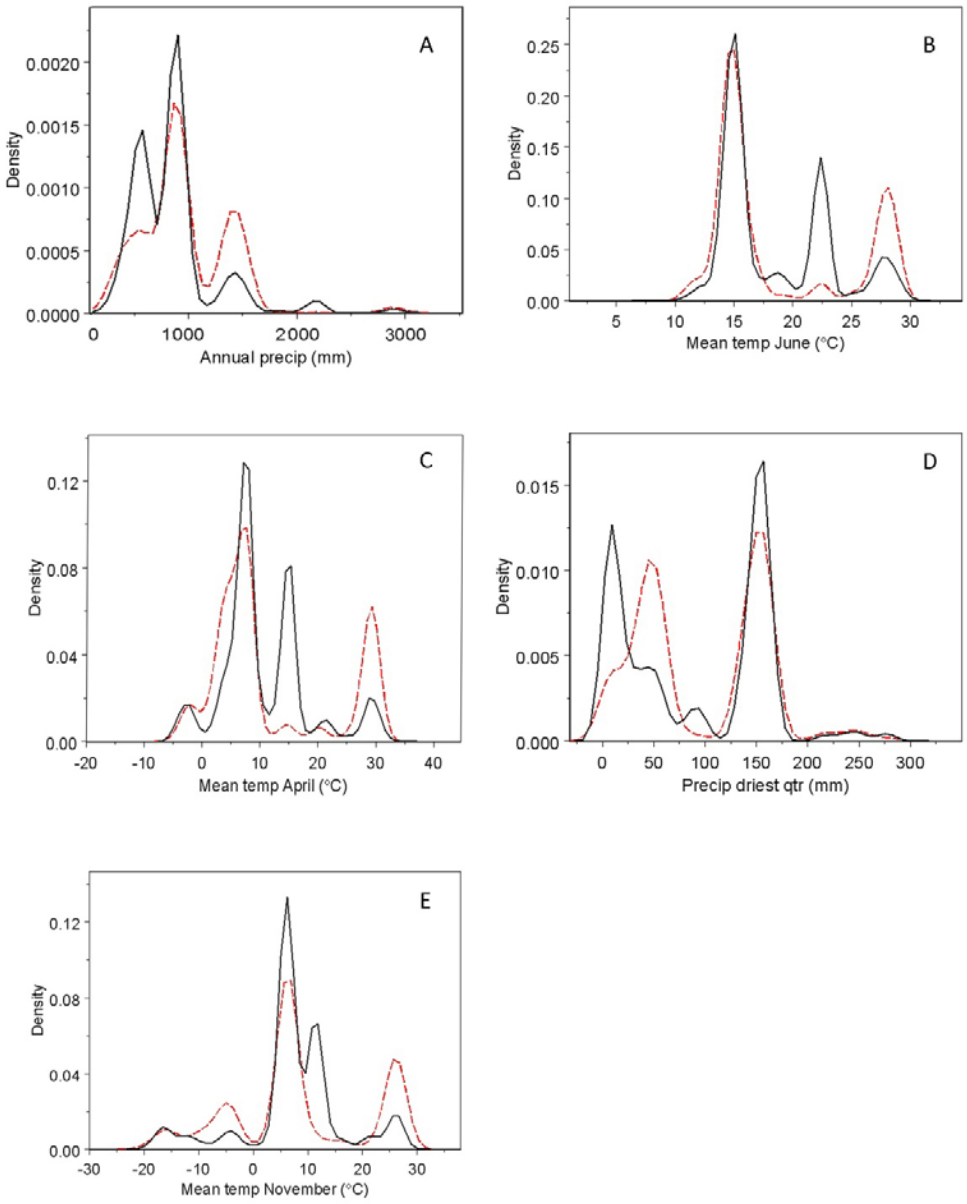

Figure 3.1. Histogram density plots for important variables. Annual Precipiation (A; mm), Mean temperature in June (B; °C), Mean temperature in April (C; °C), Precipitation of driest quarter (D; mm), and Mean temperature in November (E; °C) were the highest contributors to model accuracy as calculated by Random Forests. Density lines for cases that tested positive for avian influenza virus (AIV) are represented by dashed lines, AIV-negative by solid lines. Peaks in AIV-positive cases that exceed AIV-negative cases indicate a range of values for this predictor variable that are correlated with AIV-positivity.

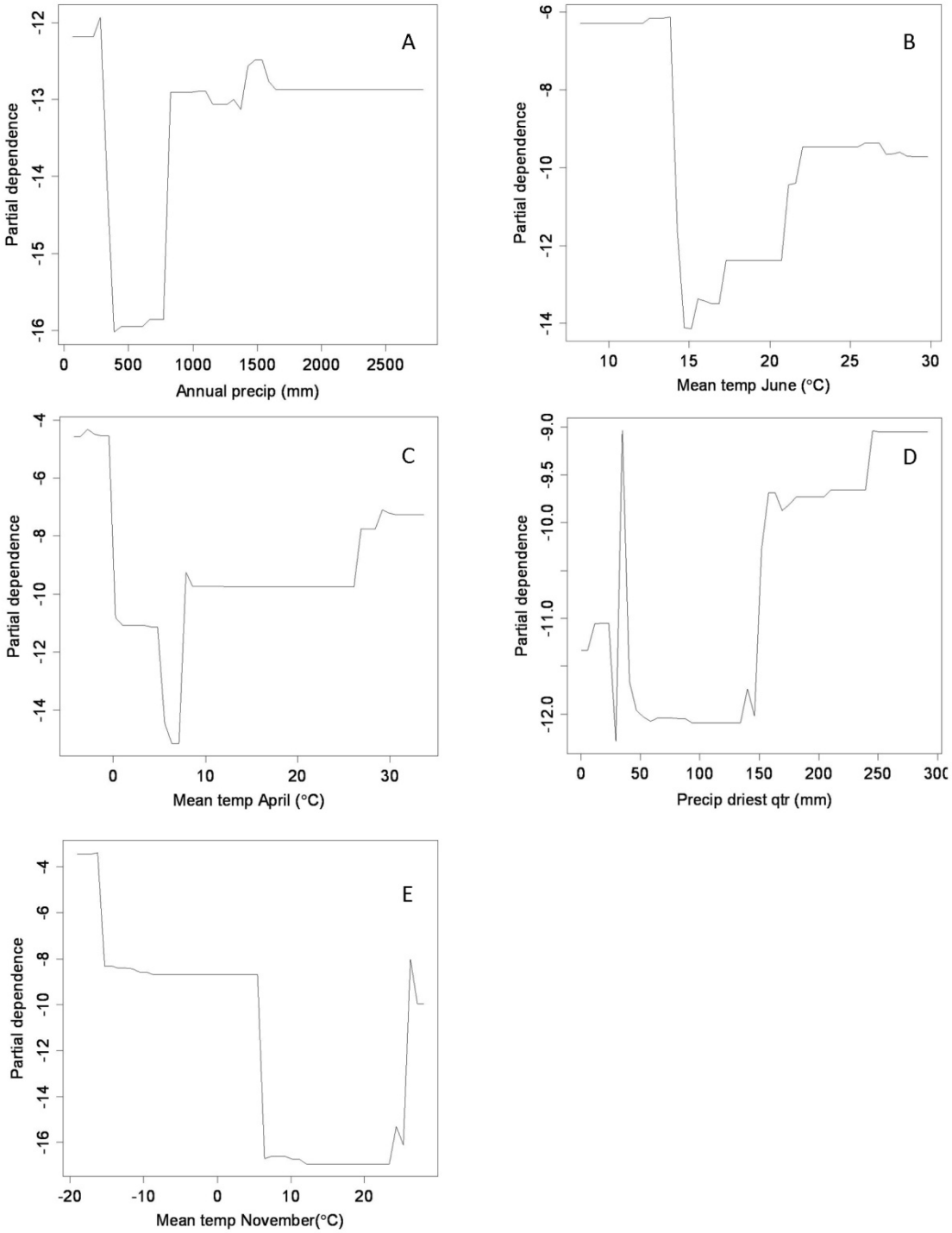

Figure 3.2. Partial dependence plots for important variables. Annual Precipiation (A; mm), Mean temperature in June (B; °C), Mean temperature in April (C; °C), Precipitation of driest quarter (D; mm), and Mean temperature in November (E; °C) were the highest contributors to model accuracy as calculated by Random Forests. Plots show the partial

dependence of avian influenza virus (AIV)-positivity on each predictor variable over the averaged effects of all other predictors. Partial dependence is best interpreted as an index of the importance of values within a variable's range and is best understood by examining general patterns in relation to the values of the predictor variable rather than the specific values of partial dependence. For example, in A, annual precipitation below 500 mm is most correlated with AIV-positivity; ranges between 500 and 1000 mm show low correlation; and ranges above 1000 mm show moderate correlation.

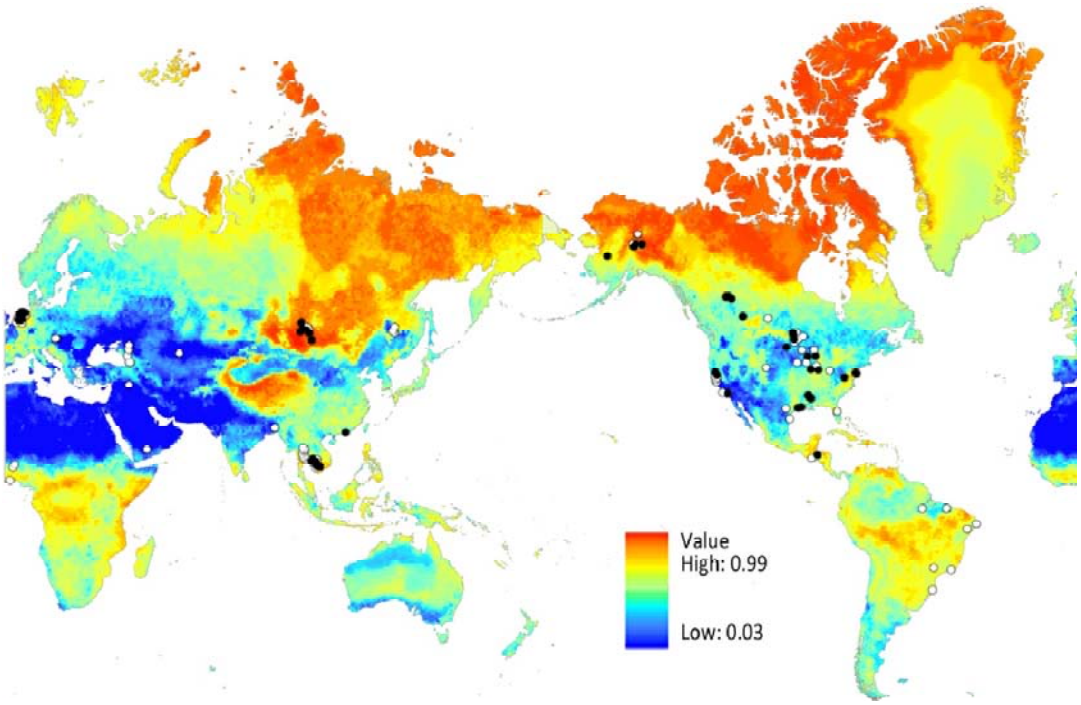

Figure 3.3. Map of predicted relative occurence index of avian influenza virus (AIV) in wild birds and surveillance locations. The map colors indicate areas of high (red) predicted occurrence of AIV to low predicted occurrence (blue). Collection locations are indicated by dots and represent the original dataset before it was divided into training and testing pools. Dots may represent more than one sample collected at those coordinates. Black dots represent locations where at least one AIV-positive sample was collected; the white dots are locations where only AIV-negative samples were collected.

LITERATURE CITED

[1] Garten RJ, Davis CT, Russell CA, Shu B, Lindstrom S, Balish A: **Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans** *Science* 2009, **325**:197-201.

[2] Stallknecht DE, Brown JD: **Wild birds and the epidemiology of avian influenza** *J Wildl Dis* 2007, **43**:15-20.

[3] Subbarao K, Klimov A, Katz J, Regnery H, Lim W, Hall H: **Characterization of an avian influenza A (H5N1) virus isolated from a child with a fatal respiratory illness** *Science* 1998, **279**:393-396.

[4] Kandun IN, Wibisono H, Sedyaningsih ER, Yusharmen DPH, Hadisoedarsuno W, Purba W: **Three Indonesian clusters of H5N1 virus infection in 2005** *N Engl J Med* 2006, **355**:2186-2194.

[5] Lignon BL: **Avian influenza virus H5N1: a review of its history and information regarding its potential to cause the next pandemic** *Semin Pediatr Infect Dis* 2005, **16**:26-35.

[6] Peiris JSM, de Jong MD, Guan Y: **Avian influenza virus (H5N1): a threat to human health** *Clin Microbiol Rev* 2007, **20**:243-267.

[7] Centers for Disease Control: Avian influenza A virus infections of humans [http://www.cdc.gov/flu/avian/gen-info/avian-flu-humans.htm].

[8] Suarez D: **Evolution of avian influenza viruses** *Vet Microbiol* 2000, **74**:15-27.

[9] Munster VJ, Baas C, Lexmond P, Waldenstrom J, Wallensten A, Fransson T: **Spatial, temporal, and species variation in prevalence of influenza A viruses in wild migratory birds** *PLoS Path* 2007, **3**:630-638.

[10] Krauss S, Walker D, Pryor P, Niles L, Chenghong L, Hinshaw VS, Webster RG: **Influenza A viruses of migrating wild aquatic birds in North America** *Vector Borne Zoonotic Dis* 2004, **4**:177-189.

[11] Webster R, Bean W, Gorman O, Chambers T, Kawaoka Y: **Evolution and ecology of influenza A viruses** *Microbiol Rev* 1992, **56**:152-179.

[12] Parmley J, Lair S, Leighton FA: **Canada's inter-agency wild bird influenza survey** *Integr Zool* 2009, **4**:409-417.

[13] CEIRS: "Centers of Excellence for Influenza Research and Surveillance (CEIRS)". National Institute of Allergy and Infectious Diseases. [http://www3.niaid.nih.gov/LabsAndResources/resources/ceirs/introduction.htm].

[14] Peterson AT: **Ecologic niche modeling and spatial patterns of disease transmission** *Emerg Infect Dis* 2006, **12**:1822-1826.

[15] Peterson AT, Sanchez-Cordero V, Beard CB, Ramsey JM: **Ecological niche modeling and potential reservoirs for Chagas disease, Mexico** *Emerg Infect Dis* 2002, **8**:662-667.

[16] Moffett A, Shackelford N, Sarkar S: **Malaria in Africa: vector species' niche models and relative risk maps** *PLoS ONE* 2007, **2**:e824.

[17] Peterson AT, Vieglais DA, Andreasen JK: **Migratory birds modeled as critical transport agents for West Nile Virus in North America** *Vector Borne Zoonotic Dis* 2003, **3**:27-37.

[18] Mak S, Morshed M, Henry B: **Ecological niche modeling of Lyme Disease in British Columbia, Canada** *J Med Entomol* 2010, **47**:99-105.

[19] Adhikari D, Chettri A, Barik SK: **Modelling the ecology and distribution of highly pathogenic avian influenza (H5N1) in the Indian subcontinent** *Curr Sci* 2009, **97**:72-78.

[20] Pfeiffer DU, Minh PQ, Martin V, Epprecht M, Otte MJ: **An analysis of the spatial and temporal patterns of highly pathogenic avian influenza occurrence in Vietnam using national surveillance data** *Vet J* 2007, **174**:302-309.

[21] Martin V, Pfeiffer DU, Zhou X, Xiao X, Prosser DJ, Guo F, Gilbert M: **Spatial distribution and risk factors of highly pathogenic avian influenza (HPAI) H5N1 in China** *PLoS Path* 2011, **7**:e1001308.

[22] Fuller TL, Saatchi SS, Curd EE, Toffelmeier E, Thomassen HA, Buermann W, Smith TB: **Mapping the risk of avian influenza in wild birds in the U.S.** *BMC Infect Dis* 2010, **10**:187.

[23] Herrick KA, Huettmann F, Runstadler J, Chernetsov N, Antonov A, Valchuk O, Gerasimov Y, Matsyna E, Matsyna A, Markovets M, Druzyaka A, Saito K: **Predictive RISK modeling of avian influenza in the Pacific Rim and beyond.** In *Risk Models and Applications, 2010*. Edited by Kremers H, Susini A. Berlin: CODATA Germany: Lecture Notes in Information Sciences; 2010:135-148.

[24] Hedenström A: **Extreme endurance migration: what is the limit to non-stop flight** *PLoS Biol* 2010, **8**:e1000362.

[25] Squires B, Macken C, Garcia-Sastre A, Godbole S, Noronha J, Hunt V, Chang R, Larsen CN, Klem E, Biersack K, Scheuermann RH: **BioHealthBase: informatics support in the elucidation of influenza virus host–pathogen interactions and virulence** *Nucleic Acids Res* 2008, **36**:D497-D503.

[26] Sanderson EW, Jaiteh M, Levy MA, Redford KH, Wannebo AV, Woolmer G: **The Human Footprint and the Last of the Wild** *Bioscience* 2002, **52**:891-904.

[27] FAO GeoNetwork: Predicted global poultry density (2005) [http://www.fao.org/geonetwork/srv/en/metadata.show?id=12719&currTab=distribution].

[28] FAO GeoNetwork: Predicted global pig density (2005) [http://www.fao.org/geonetwork/srv/en/metadata.show?id=12719&currTab=distribution].

[29] Beyer HL: Geospatial Modelling Environment [http://www.spatialecology.com/].

[30] Breiman L: **Random Forests** *Mach Learn* 2001, **45**:5-32.

[31] Prasad AM, Iverson LR, Liaw A: **Newer classification and regression tree techniques: bagging and Random Forests for ecological prediction** *Ecosystems* 2006, **9**:181-199.

[32] Cutler DR, Edwards TC, Jr., Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ: **Random Forests for classification in ecology** *Ecology* 2007, **88**:2783-2792.

[33] Liaw A, Wiener M: **Classification and regression by RandomForest** *R News* 2002, **2**:18-22.

[34] R Development Core Team: **R: A language and environment for statistical computing**. Vienna: R Foundation for Statistical Computing; 2010.

[35] Friedman JH: **Greedy function approximation: a gradient boosting machine** *Ann Stat* 2001, **29**:1189-1232.

[36] Hegel TM, Cushman SA, Evans J, Huettmann F: **Current state of the art for statistical modelling of species distributions.** In *Spatial Complexity, Informatics, and Wildlife Conservation*. Edited by Cushman SA, Huettmann F. Tokyo: Springer; 2010:273-311.

[37] Schröder B: **ROC-Plotting and AUC Calculation Transferability Test v 1.3-7**. Potsdam: Universität Potsdam; 2006.

[38] Hosmer DW, Lemeshow S: *Applied Logistic Regression*. New York: Wiley; 2000.

[39] Lowen AC, Mubareka S, Steel J, Palese P: **Influenza virus transmission is dependent on relative humidity and temperature** *PLoS Path* 2007, **3**:e151.

[40] Wood J, Choi Y, Chappie D, Rogers J, Kaye I: **Environmental persistence of a highly pathogenic avian influenza (H5N1) virus** *Environmental Science and Technology* 2010, **44**:7515-7520.

[41] Brown JD, Goekjian G, Poulson R, Valeika S, Stallknecht DE: **Avian influenza virus in water: infectivity is dependent on pH, salinity and temperature** *Vet Microbiol* 2007, **136**:20-26.

[42] Webster RG, Yakhno M, Hinshaw VS, Bean WJ, Murti KG: **Intestinal influenza: replication and characterization of influenza viruses in ducks** *Virology* 1978, **84**:268-278.

[43] Löndt BZ, Nunez A, Banks J, Nili H, Johnson LK, Alexander DJ: **Pathogenesis of highly pathogenic avian influenza A/turkey/Turkey/1/2005 H5N1 in Pekin ducks (Anas platyrhynchos) infected experimentally** *Avian Pathol* 2008, **37**:619-627.

[44] Gilbert M, Xiao X, Pfeiffer DU, Spprecht M, Boles S, Czarnecki C: **Mapping H5N1 highly pathogenic avian influenza risk in Southeast Asia** *Proc Natl Acad Sci USA* 2008, **105**:4769-4774.

[45] Hijmans RJ, Cameron SE, Parra JL, Jones P, Jarvis A: **Very high resolution interpolated climate surfaces for global land areas** *Int J Climatol* 2005, **25**:1965-1978.

[46] Center for International Earth Science Information Network (CIESIN), Columbia University, Centro Internacional de Agricultura Tropical (CIAT): Gridded Population of the World Version 3 (GPWv3): Population Density Grids [http://sedac.ciesin.columbia.edu/gpw].

[47] Robinson TP, Franceschini G, Wint G: **The Food and Agriculture Organization's Gridded Livestock of the World** *Vet Ital* 2007, **43**:745-751.

[48] Lehner B, Döll P: **Development and validation of a global database of lakes, reservoirs and wetlands** *J Hydrol* 2004, **296**:1-22.

CHAPTER 4

Modeling avian influenza with Random Forests: under-sampling and model selection for unbalanced prevalence in surveillance data[3]

ABSTRACT

Highly imbalanced prevalence is a common occurrence in wildlife surveillance data and may lead to poor performance in predictive models. The percentage of AIV-infected birds in a dataset is often less than 1%. To address imbalanced prevalence in data such as these, we evaluated the effects of a balancing algorithm, a model selection algorithm, a combination of balancing and model selection, and an under-sampling method on the prediction accuracy of classification models constructed using Random Forests. We developed an ecological niche model for low-pathogenicity avian influenza in wild birds using data from two independent AIV surveillance projects, and then applied the findings to create a predictive map of a study area around the Pacific Rim. Although the best performing model was a combination of balancing and model selection, the balancing algorithm boosted accuracy more significantly than other methods when applied to datasets with low prevalence (< 5%). When implemented through R, this balancing algorithm is a relatively simple method for improving signal. In addition, repeated sampling in a single location appears to be more detrimental to data used for machine-learning purposes than imbalanced prevalence.

---

[3]Herrick, K.A., Heuttmann, F. Modeling avian influenza with Random Forests: under-sampling and model selection for unbalanced prevalence in surveillance data. Ecological Informatics *in preparation.*

1. INTRODUCTION

Although the highly pathogenic avian influenza (HPAI) strain H5N1 has dominated modeling and spatiotemporal analysis efforts, it is just one of many strains of avian influenza virus (AIV). HPAI H5N1 isolation from wild birds, even in areas of high risk or active outbreak, is very low (<1%) (Chen et al., 2006; Globig et al., 2009; Krauss et al., 2007), whereas low-pathogenicity avian influenza (LPAI) is nearly ubiquitous in wild bird populations. The prevalence of LPAI can range from 60% to less than 1% depending on the species sampled, season, and location of samplings. Table 4.1 illustrates this range of prevalence from selected surveillance projects. While wild waterfowl are thought to be the reservoir for AIV, AIV can also infect a wide range of both domestic and wild birds including Psittaciformes (Kaleta et al., 2007), Falconiformes (Goyal et al., 2010; Hall et al., 2009), Passerines (Fuller et al., 2010), ratites (Abolnik et al., 2010), and Galliformes such as chickens, quail, and turkeys (Perkins and Swayne, 2001). It should be noted that some of these examples are experimental infections and unlikely to reflect viral endemism in the species. In the wild, surveillance studies found that Anseriformes and Charadriiformes were the most commonly infected (Munster et al., 2007; Stallknecht and Shane, 1988). AIV in wild birds is a crucial component in the dynamics of transmission and risk factors associated with HPAI H5N1 or any other strain of AIV.

HPAI strains likely arise from LPAI strains and a better understanding of LPAI could contribute to improved understanding and prediction of HPAI strains. H5N1 is not the only strain of AIV that has demonstrated the potential to become highly pathogenic. Under experimental conditions, avirulent strains, which cause no clinical symptoms and

replicate poorly in chickens, grow increasingly virulent and convert to HPAI simply by being passed from bird to bird (Cilloni et al., 2010; Ito et al., 2001). While HPAI H5N1 has yet to be identified in the Western Hemisphere, there have been five outbreaks of other HPAI subtypes: an H5N2 outbreak occurred in Pennsylvania in 1983 (Davison et al., 2003); in 2002, Chile experienced an H7N3 outbreak (Senne, 2007); an H7N2 outbreak occurred in New York in 2003 (Senne, 2007); and in 2004, an H5N2 outbreak was documented in Texas (Lee et al., 2005; Pelzel et al., 2006) and an H7N3 outbreak in Saskatchewan, Canada (Pasick et al., 2010).

Few studies have examined the risk factors and prediction of AIV in wild bird populations and those that have focused on HPAI H5N1. A study by Kilpatrick et al. (2006) included wild bird data in the calculations of risk factors predicting the spread of HPAI H5N1. Spatial analyses were conducted on the spread of HPAI H5N1 in wild migratory waterfowl in Russia (Gilbert et al., 2006; Reemers et al., 2010). Fuller et al. (2010) constructed a model for wild birds in the continental United States to identify important predictors for LPAI outbreak. It has also been suggested that identifying specific bioclimatic and landcover characteristics of regions in which HPAI H5N1 outbreaks occur would improve the prediction of pathogen dispersal (Sengupta et al., 2007). Sengupta's study identifies terrestrial ecoregions (Olson et al., 2001) in which a disproportionate number of AIV cases were reported. This study was criticized because ecoregions are not homogeneous in terms of climate or land-use, and the model focused on wild birds rather than domestic poultry (Jourdain et al., 2007). However, to fully understand disease and transmission, identifying the ecological context in which

outbreaks occur is crucial. There is great value in "identifying specific climatic and vegetation zones that are important in the life cycle of *Anatidae*" (Sengupta et al., 2007). We agree that quantifying the aspatial conditions in areas of importance could help to understand correlates of AIV-positivity. Rather than using ecoregions, this is better achieved by defining an environmental niche for AIV-positivity. This description is then extrapolated to statistically similar niches, and then visualized in geographic space. With the exception of prior publications by this group (Chapter 3, this volume; Herrick et al., 2013), as of this writing, two other models for strains other than HPAI include a logistic regression analysis of wild birds in the continental United States (Fuller et al., 2010) and a risk map built with Maximum Entropy Modeling that included wild birds and domestic poultry ((Moriguchi et al., 2012). The Fuller model included all available AIV isolates and did not specifically test for HPAI while the Moriguchi model included both HPAI and LPAI strains.

Studies of HPAI H5N1 risk factors are primarily spatiotemporal analyses (Cecchi et al., 2008; Kilpatrick et al., 2006; Oyana et al., 2006; Reemers et al., 2010; Ward et al., 2008). A few studies employed logistic regression techniques (Gilbert et al., 2008; Pfeiffer et al., 2007; Reemers et al., 2010) to analyze risk factors. Ecological niche models were constructed using the Genetic Algorithm for Rule Set Production (GARP) (Adhikari et al., 2009; Williams et al., 2008). These studies consistently identified a number of risk factors for HPAI H5N1 including 1) agricultural land use, particularly rice cropping (Gilbert et al., 2008; Paul et al., 2010), floodplain agriculture (Cecchi et al., 2008), or aquaculture (Gilbert et al., 2008); 2) domestic poultry, while the likelihood of

contracting H5N1 showed some dependence upon the type of chicken (i.e. layer vs. broiler) (Thomas et al., 2005), the presence of domestic waterfowl (free-grazing ducks in particular) was a more common factor (Gilbert et al., 2006; Gilbert et al., 2008; Hogerwerf et al., 2010; Martin et al., 2011; Paul et al., 2010; Pfeiffer et al., 2007) high human population density and associated anthropogenic factors such as the proximity of roads and cities (Adhikari et al., 2009; Gilbert et al., 2006; Gilbert et al., 2008; Hogerwerf et al., 2010; Martin et al., 2011; Ward et al., 2008); 4) topographical factors including low elevation (Gilbert et al., 2006; Gilbert et al., 2008; Martin et al., 2011), low slope angle, and seasonal flooding (Ward et al., 2008); and 5) the influence of migratory waterfowl (Cecchi et al., 2008; Kilpatrick et al., 2006; Oyana et al., 2006).

Due to the complexity of the underlying data, identification of risk factors and environmental characteristics associated with AIV, predicting the occurrence of AIV-positivity is best addressed by constructing ecological niche models using data-mining machine learning methods (Hegel et al., 2010). We chose the machine-learning algorithm Random Forests to identify important predictor variables and generate a model of relative predicted occurrence. Random Forests is an ensemble method that aggregates the findings of multiple decision trees to improve accuracy (Seni and Elder, 2010). One important feature is that it produces a list of predictor variables ranked by contribution to model accuracy. Random Forests is considered generally immune to overfitting (Breiman, 2001) and performs with comparable accuracy to logistic regression on unbalanced datasets (Ruiz-Gazen and Villa, 2007). When implemented through R 2.11.1 Statistical Programming Language (R Development Core Team, 2010) using the

randomForest package (Liaw and Wiener, 2002), it is free to use and the code is customizable.

In addition, we were interested in addressing the problem of highly imbalanced data from AIV surveillance in wild birds. As the application of quantitative modeling expands, attempts to learn from highly imbalanced datasets may pose problems when the minority class, which is usually the class of interest, is much smaller than the majority class. Imbalanced prevalence is common in wildlife surveillance data and AIV in wild birds is no exception. Imbalanced datasets can reduce the performance of machine-learning algorithms and is a topic of active interest in the Machine Learning community (He and Garcia, 2009). Users run the risk of attempting to train a model on a subset containing no instances of the minority class (Chen *et al*., 2004). While the randomForest package for R lacks the weighting function of the original Random Forests algorithm developed by Breiman (2001), Chen *et al.* (2004) demonstrated that there was "no clear winner" between balancing and weighting. Both were effective when using Random Forests to classify highly imbalanced data (prevalence of minority class < 4%). In particular, repeated random sub-sampling is a relatively simple balancing algorithm that improves accuracy by ensuring each training set has an equal number of positive and negative instances. This algorithm has proven to be effective at balancing data for use in Random Forests when predicting tree species distribution (Freeman et al., 2012), predicting disease (Khalilia et al., 2011), and mapping HPAI H5N1 risk using a multiple logistic regression model (Gilbert et al., 2008). In order to further improve model accuracy, we implemented a model selection algorithm that determines the most

parsimonious subset of predictor variables. Although the relative ranking of important variables by Random Forests is not sensitive to either noise or highly correlated variables (Genuer et al., 2010), previous studies have used model selection to improve their prediction accuracy (Evans and Cushman, 2009; Murphy et al., 2010). We used a process that first ranked predictor variables by their contribution to model accuracy and then, through stepwise, descending elimination, chose the most parsimonious subset of predictor variables that produced the smallest estimated error.

2. MATERIALS AND METHODS

2.1 Predictor variables

Thirty-eight geographic, anthropogenic, and bioclimatic variables were used as predictors in this project (Table 4.2). Because previous studies have correlated low elevation and slope with HPAI H5N1 outbreak (Adhikari et al., 2009), we included elevation and slope as predictor variables. Slope (m) was calculated from elevation using the ArcGIS Spatial Analyst tool. Rather than use ecoregions (Sengupta et al., 2007), we included landcover categories, which describe vegetation and land use at a higher resolution. These categories were primarily represented as vegetation type, but also included water cover, cropland, and artificial surfaces (Bartholomé and Belward, 2005). We condensed the original categories into 12 more general ones because not all the original categories were represented in the training data. Due to the predominance of waterfowl collected in the surveillance efforts, we calculated the Euclidean distance from rivers and lakes based on a layer composed of combined hydrologic freshwater features (Lehner and Döll, 2004)

using ArcGIS Spatial Analyst tools. As shorebirds appear to be the second most common carrier of AIV (Stallknecht and Shane, 1988), we calculated the Euclidean distance from coastline based on a coastline data layer included with ArcMap. Anthropogenic factors were represented by human population density (Center for International Earth Science Information Network (CIESIN), et al., 2005) and the Human Influence Index (Sanderson et al., 2002) calculated from human presence indicators. Bioclimatic predictor variables included mean temperature for each month as well as 19 other temperature and precipitation measures such as mean annual temperature, temperature seasonality (annual range in temperature), and precipitation of driest and wettest quarters. All predictors used for these analyses are presented in detail in Table 4.2.

2.2 Wild bird data

We used two datasets (the A3IR and CIWBI databases) from independent wild bird AIV surveillance projects and two manipulations of these databases (ALL and UNIQUE). Descriptions for these databases are summarized in Table 4.3. Both A3IR and CIWBI were similar in that each data point represents an individual bird sample with latitude/longitude coordinates of sampling location, species, sampling date, and AIV-positive status. In both projects, AIV-positivity was determined by RRT-PCR viral matrix screening using the same protocol for detection (Spackman et al., 2002).

A3IR: This database contained wild bird data collected in Alaska between 2005 – 2010 by the Alaska Asia Avian Influenza Research group (A3IR). A study using portions of these data have been published previously (Herrick et al., 2010). Since this previous publication, additional samples were collected, the samples collected outside of North

America were excluded, and the database itself was curated and migrated to a new platform. Samples were taken by cloacal and oropharyngeal swabs primarily from waterfowl and shorebirds, although other orders such as Passeriformes, Falconiformes, and Pelicaniformes were also included. Environmental and fecal swabs were excluded from analysis as the source species could not be positively identified. Likewise, samples where the species was listed as "unknown" were also excluded from analysis. The final database contains 40,741 samples, 4% of which were AIV-positive.

CIWBI: Starting in 2005, Canada's Inter-agency Wild Bird Influenza survey (CIWBI) has conducted yearly surveys of live wild birds and wild birds found dead (Parmley et al., 2011; Parmley et al., 2009). The live survey collected cloacal and oropharyngeal swabs from waterfowl and shorebirds; the dead bird survey included all species submitted to regional laboratories for AIV testing. These data are generated by Canada's inter-agency Wild Bird Influenza Survey and made available by the Canadian Cooperative Wildlife Health Centre. Data are available for download at [http://www.ccwhc.ca/aiv_bird_surveys_2011.php]. We used both live and dead bird data collected in 2010 – 2011. While AIV may have an effect on wild bird mortality, only 0.5% of the dead bird samples were AIV-positive, while 25% of the live bird samples were AIV-positive. Both these percentages are within the range reported by other surveillance studies (Table 4.1) and it does not appear that the inclusion of 1757 dead bird samples strongly skewed the combined dataset. This dataset contained 4,298 samples, 15% of which were AIV-positive.

The A3IR and CIWBI databases were manipulated as follows. ALL: the A3IR and CIWBI databases were combined for a total of 45,039 samples, with 2,245 AIV-positive cases (5%). UNIQUE: duplicate data points with identical latitude/longitude coordinates from the combined databases were excluded from analysis. A latitude/longitude combination was considered AIV-positive if any of the duplicates were positive; otherwise, the point was negative if all duplicates were AIV-negative. The resulting database contained 1,930 unique latitude/longitude combinations, 10% of which were AIV-positive. Of these, A3IR and CIWBI contributed 1,037 and 893 locations, respectively.

2.3 Random Forests, balancing, and model selection

Predictive models were developed from classification trees built using the Random Forests algorithm (Breiman, 2001; Cutler et al., 2007) integrated with GIS. Random Forests is a data-mining method based on building an ensemble of classification decision trees. Each tree is constructed from a bootstrap sample and split at each node by the best predictor from a very small, randomly chosen subset of the predictor variable pool. The two parameters that can be specified by the user are the total number of trees in the forest and the number of predictors to try at each node. Random Forests are not particularly sensitive to these parameters (Liaw and Wiener, 2002) and the optimal range is very wide (Breiman, 2001). One advantage of Random Forests is that it calculates an internal estimate of error: approximately 30% of the cases are randomly set aside or kept "out-of-bag" (OOB) and not used for training. When a tree is grown, each OOB case is run down the tree and the proportion of times cases are misclassified, averaged over all cases is the

OOB error estimate. Random Forests have demonstrated high classification accuracy and usefulness in ecological settings (Craig and Huettmann, 2008; Evans and Cushman, 2009; Prasad et al., 2006).

Models were constructed by applying four different methods to the four databases in R using the randomForest package (Liaw and Wiener, 2002). The methods used were 1) default, 2) repeated random sub-sampling (RSS), 3) model improvement ratio (MIR), and 4) MIR combined with RSS (MIR+RSS). We constructed ten replicates of each model. The default method involved constructing a Random Forests model using default settings where the number of trees was 500 and the number of predictors to try at each node (mtry) was the square root of the number of predictors (Cutler et al., 2007). The MIR algorithm was based on previous work done by Evans and Cushman (2009) and Murphy et al. (2010). Although Random Forests is relatively immune to overfitting (Breiman, 2001), model selection, or the removal of non-contributing predictor variables, can increase accuracy, decrease noise, and decrease OOB error. MIR selects the model with the optimal set of predictor variables based on the lowest total OOB and lowest maximum within-class error. First, a Random Forests model is constructed using the default settings with all predictor variables. Next, the model is re-run using a subset of predictor variables with importance values greater than or equal to a specified threshold value. The threshold value begins at 0.1 and is increased by 0.2 to a maximum of 0.9. All models are compared and the subset of predictor variables contributing to the model with the lowest OOB and maximum within-class error are used in a final Random Forests model. Repeated random sub-sampling (RSS) has been used in previous models

(Freeman et al., 2012; Gilbert et al., 2008), but is best described by Khalilia et al. (2011). This method was designed to up-weight the minority class (in our case the AIV-positive samples) by creating a subset containing all the positive samples and an equal number of randomly selected negative samples. Negative samples were re-drawn with replacement for each subsequent subset. The default Random Forests algorithm built a forest from a specified number of trees (500, in our case). We built a single tree from a balanced subset, drew a new balanced subset, and repeated this process 500 times. This forest was then aggregated into a single randomForest object using the "combine" command for further analysis. MIR+RSS was identical to the MIR algorithm with the addition of repeated random sub-sampling in the threshold testing steps. The model constructed at each threshold ratio was an aggregated forest of single trees built from random balanced subsets. In this way, our highly imbalanced dataset was up-weighted and combined with model selection to eliminate non-contributing predictor variables, reduce noise, and enhance model accuracy. Salford Predictive Modeler (Salford Systems) is data-mining software unrelated to R that allows the user to implement Random Forests. The benefits of this software over the R version are that it continues to undergo research and improvement under one of the algorithm's original co-authors, it runs under a convenient GUI, and it produces a number of descriptive results and graphics not available in the R version. The deficits are that it requires the purchase of a license, and it lacks some of the features of the randomForest R package, most notably the ability to produce partial dependence plots. Salford Random Forests was applied to the four datasets using the following settings: classification, Class Weights = Balanced, Number of predictors at

each node = 6, number of trees = 500. The results of a single forest for each dataset are presented as a comparison to the R methods.

2.4 Predictive map

Each of the Random Forests models constructed in the previous step was applied to a new set of points that formed a regularly spaced grid covering the study area. Each grid point was given a predicted AIV-positive value between 0-1 based on the model applied. These values were then interpolated between points using the Inverse Distance Weighted (IDW) tool in ArcGIS, resulting in a map of predicted occurrence of AIV.

2.5 Statistical analyses

To evaluate model performance, we first compared the experimental AIV-status of the original points to their predicted values from the predicted occurrence map and then plotted a Receiver Operating Characteristic (ROC) curve. To produce an ROC, the true-positive rates are plotted against their false-positive rates. The resulting Area Under the Curve (AUC) serves as a convenient summary measure with which to compare model performance. ROC and AUC have demonstrated their reliability and usefulness in evaluating ecological models and predictive models in machine-learning and data-mining (Fielding and Bell, 1997; Zhou et al., 2007). ROC curves and AUC values were calculated using the ROCR package in R (Sing et al., 2005).

For each dataset and each algorithm, we constructed 10 replicates and calculated their AUCs. In order to determine the probability that mean AUCs were significantly different from an AUC of 0.5 produced by a pure random model, we calculated upper and lower confidence bounds, and p-values based on t-distributions using Microsoft Excel.

Using the null hypothesis that there were no differences between group AUC means, we conducted a one-way omnibus ANOVA in R. Databases were divided into groups by method, and two-way pairwise t-tests were conducted on the AUCs in R. Finally, the combination of best model and database were selected based on AUC values that met or exceeded the critical value of 0.7, which is the conventional AUC threshold of acceptable strength (Hosmer and Lemeshow, 2000).

2.6 Variable importance

The best performing model was chosen by its AUC, and its attributes were further analyzed. Random Forests returns a list of predictor variables ranked in order of contribution to model accuracy. The two main ranking criteria it uses are Mean Decrease in Gini (MDG), which is the average decrease in Gini impurity at each node, and Mean Decrease in Accuracy (MDA), which is calculated as the mean accuracy for a predictor minus the decrease in model accuracy when this predictor is randomly permuted. We chose to rank our predictor variables' importance by MDA, which has demonstrated better robustness and stability than MDG (Nicodemus, 2011). We ranked predictors based on their mean MDA scores over ten replicates and normalized the list to the highest scoring predictor.

To examine the contributions of each important predictor variable, we produced partial dependence plots using the randomForest package in R. Partial dependence is used to summarize the influence of a predictor variable above the average noise of other predictors (Friedman, 2001). The partial dependence of a predictor variable can be examined by observing general trends over the variable's range rather than specific

values of partial dependence. We created density plots in Spotfire S+ (TIBCO, v.8.2) to visualize the relative probability of AIV-positive versus AIV-negative samples across a predictor variable's range of values. Peaks in density plots may help to elucidate important ranges in a predictor variable's values that correspond with AIV-positivity, and is one method of comparing characteristics of AIV-negative points to AIV-positive points. R code for the algorithms and partial dependence plots are presented as Appendix G.

2.7 Cross-model comparisons

To determine the generalizability of the predictions, we applied the highest performing model from each database group to the other databases. The predicted relative occurrence map produced by the best performing model for each database was compared to the actual AIV-status of each bird. AUCs were calculated by plotting the experimental-positive rates against their false-positive rates.

2.8 Research design

Each method (Default, MIR, RSS, MIR+RSS) was applied to each database (A3IR, CIWBI, ALL, UNIQUE) and ten replicate models were constructed for each combination. Table 4.4 summarizes each of the experimental methods: Default, MIR, RSS, MIR + RSS, and Salford. Each model was then applied to a grid of regularly-spaced points and the predicted AIV-positive values were interpolated between points across the study area to create a predictive map for each model. AUC values for each model were calculated based on these maps and the mean of the ten replicates was calculated for each database/method combination. The highest scoring mean AUC for each database

determined the database/method combination that was then used for the cross-model comparison. The points from each database were compared with the ten replicates of the predictive maps from these combinations to generate AUC values. Finally, the highest scoring database/method combination from the first step was used to construct a predictive map, and its important predictor variables were further analyzed. A diagram outlining these steps is presented in Figure 4.1.

3. RESULTS

In order to develop a powerful predictive model for highly unbalanced datasets, we tested four different variations of the Random Forests ensemble data-mining algorithm on two different databases and two manipulations of these databases. We based model accuracy on AUC values and presented a predictive map and analyses of important predictor variables based on the highest performing combination.

3.1. Model Performance

ROC curves for each method, except for Salford, were plotted together by database (Fig. 4.2). In addition to providing a visual comparison, ROC curves also represent model performance characteristics. For instance, for the A3IR database, Default, MIR, and MIR+RSS models produced ROC curves indistinguishable from pure random, which would be a straight, diagonal, line (Fig. 4.2A). RSS outperformed the other models and produced a balanced, bow-shaped curve. The CIWBI models have a different shape that produces a strong peak in the lower end of the false-positive range (Fig. 4.2B). This implies that the model is “conservative” in that it has a low rate of false positives, but

does this at the expense of some true positives (Fawcett, 2004). The UNIQUE models, in contrast, show high values in the upper right hand quadrant meaning they detect more true positives, but at an increased false-positive rate (Fig. 4.2D). It appears the nature of the database, rather than balancing or model selection, influences the shape of the curve, which was consistent within each database.

To determine model performance over completely random assignment of class (AUC=0.5), we calculated 95% confidence intervals for each database/method based on t-distributions and p-values (Table 4.5). Salford AUC results are presented here as comparison. All models produced p-values <0.0001 except for those constructed using the A3IR database. With the exception of the RSS method applied to A3IR ($p<0.002$), all methods performed no better than random. The mean OOB error and treesize (number of nodes) are also presented in Table 4.2. Although MIR was designed to choose the model with the lowest OOB error rate, both the MIR and MIR+RSS methods regardless of database were nearly identical to the Default method.

We then categorized the databases/methods according to the Hosmer and Lemeshow (2000) threshold of 0.7 for an acceptable AUC value and presented the results in Figure 4.3. UNIQUE and CIWBI databases both produced AUCs at or above 0.7. The highest-scoring model was MIR+RSS on the UNIQUE database, which returned an AUC of 0.767 ± 0.002. For CIWBI, the MIR+RSS algorithm produced a mean AUC of 0.697 ± 0.01 (n=10), which was the highest-scoring method over Default, MIR, and RSS (0.694 ± 0.01, 0.687 ± 0.01, 0.686 ± 0.01 $p<0.05$, respectively). Random Forests were unable to detect a signal above random on the A3IR database (AUCs ranged between 0.48 ± 0.02 –

0.54±0.03). A slightly stronger signal was detected in the ALL database; however, the highest value (RSS, 0.67 ± 0.02) was still below the critical value of 0.7. RSS performed significantly better than the other methods in the A3IR and ALL databases ($p<0.005$, $n=10$), although the resulting AUCs were still below the acceptable threshold of 0.7 (0.5422 ± 0.03, 0.67 ± 0.02, respectively). Overall, the RSS method outperformed the other methods on the A3IR and CIWBI databases. The MIR+RSS method produced the highest performing models on the ALL and UNIQUE databases although the margin was very slender. Salford outperformed all R methods for each dataset based on AUC scores, however, since we were unable to handle the results with exactly the same experimental methods used for the other cases (i.e.aggregating the results, creating a randomForest object for partial dependence analysis), it was excluded from further analysis.

3.2. Cross-model comparison

To determine whether model performance was due to the database or due to the method used, we applied the highest performing model from each database group to the other databases (Fig. 4.4). The CIWBI database produced the highest mean AUC across all the models applied to it (mean AUC = 0.71; A3IR/RSS 0.628 ± 0.025; UNIQUE/MIR+RSS 0.730 ± 0.002). The CIWBI database also produced the highest overall cross-model AUC value (ALL/RSS on CIWBI; 0.774 ± 0.004), which was higher than the highest performing native model. The UNIQUE/MIR+RSS model produced the highest mean AUC across all databases (0.661) and outperformed the native model on the A3IR and CIWBI databases (0.602 ± 0.001, 0.739 ± 0.002, respectively), but not on the ALL database where it performed relatively poorly (0.535 ± 0.016). The A3IR/RSS model was

the worst performer across all databases (mean AUC = 0.561) and scored the lowest AUC on the CIWBI and ALL databases (0.628 ± 0.025, 0.602 ± 0.001, respectively), though it performed marginally better than ALL/RSS on the UNIQUE database (0.614 ± 0.025). On both the ALL and UNIQUE databases, native models outperformed the other models. In summary, all models had good prediction accuracy when applied to the CIWBI database and consistently produced high AUCs. The highest cross-model AUC value was the ALL/RSS model applied to the CIWBI database. The UNIQUE/MIR+RSS model outperformed the other models when applied to the other databases.

To further examine differences between the databases and understand factors that may contribute to the predictive models developed from them, we plotted the density of the mean temperature in December of all points (AIV-positive and AIV-negative, combined) from the A3IR database against UNIQUE (Fig. 4.5A) and CIWBI (Fig. 4.5B). The mean temperature in December was chosen because of its high variable importance score in all three of these databases. Both the CIWBI and UNIQUE databases have a fairly broad range of sampling between -30 °C and 10 °C, whereas A3IR has a very strong peak at -22.7 °C implying that most of the samples were collected over a very narrow range. Exacerbating the imbalance is the fact that the A3IR database is much larger than the other two and contains 40,740 samples whereas CIWBI contains 4,298 samples and UNIQUE contains 1,930 samples.

3.3. Variable importance

The UNIQUE/MIR+RSS model had the highest AUC score and was used for subsequent analyses. Of the predictor variables, landcover had the highest variable importance score

followed by the mean temperature in April, temperature seasonality (annual range in temperature calculated as standard deviation x 100), distance to river/lake/wetland, slope, and human population density (Table 4.2). While the most samples were collected in needleleaf landcover, more AIV-positive samples were collected in areas of broadleaf, mixed leaf, mosaic (shrubland), herbaceous, and managed landcover (Fig. 4.6A). The partial dependence of AIV-positivity was the highest on needleleaf landcover, followed by mosaic cover, and sparsely vegetated regions. Managed landcover and artificial surfaces had moderate partial dependence as did water, ice and/or snow covered surfaces (Fig. 4.7A). While the sampling range for the mean temperature in April ranged from over 10 °C down to almost -30 °C, the mode for both the AIV-positive and –negatives samples were -2.2 °C (Fig. 4.6B). The partial dependence of AIV-positivity on the mean temperature in April was highest below -5 °C, with a great deal of variability between 0-10 °C (Fig. 4.7B). Temperature seasonality ranged between approximately 2000 to over 20,000. Most samples were collected in the high end of this range with the modes for AIV-negative and –positive at 14,343 and 14,355, respectively (Fig. 4.6C). The partial dependence of temperature seasonality was moderate up to approximately this value, where it peaks (Fig. 4.7C) indicating a niche with high temperature seasonality. As the majority of samples came from dabbling ducks, most samples were collected near lakes, rivers, or wetlands and the mode for both AIV-negative and –positive samples is 0 (Fig. 4.6D). However, the partial dependence of the distance from rivers, lakes, or wetlands shows only minor peaks at low values, and then reaches its highest value above 30,000 m (Fig. 4.7D). Slope was 0° for the collection location of most samples (Fig. 4.6E);

however, slope demonstrated low partial dependence at this values, but peaked at higher values of 6° through 12° (Fig. 4.7E).Sampling mainly occurred in unpopulated areas (mode = 0 for both AIV-positive and –negative groups), but was carried out across a very wide range of human population densities, up to nearly 7,500 persons/km^2 (i.e. downtown Toronto) (Fig. 4.6F).The partial dependence on human population density is highest at a value of 0 persons/km^2, drops to a low value of partial dependence, and then gradually increases to moderate dependence at approximately 3000 persons/km^2 (Fig. 4.7F). Based on these predictors, the ecological niche for AIV-positive samples was described as having wooded landcover, cold spring temperatures below 0 °C, with highly variable temperatures. This niche is far from rivers, lakes, or wetlands, in moderately sloped areas with a very low human population density.

3.4 Predictive map

We were able to visualize predicted risk areas for LPAI in wild birds in geographic space based on the model constructed by applying the MIR+RSS method to the UNIQUE database (Fig. 4.8). The largest area of contiguous high predicted occurrence (red area) spanned East and North Asia that included northern steppe, montane tundra, and boreal forest. In Central Asia, the Kazakh steppe and Tibetan plateau are predicted to have moderately high occurrence (yellow-orange). Predicted occurrence in Alaska was similar, with high-scoring areas in boreal forest areas. Interestingly, montane tundra and other high elevation areas such as the St. Elias range, mountainous areas along the southeastern coast, and the island of Kodiak also display moderately-high to high predicted occurrence. Overall, northern regions displayed the largest areas of contiguous high

predicted occurrence. Mottled areas of high and moderately-high predicted occurrence can be seen in temperate areas such as the continental United Stated as well as Indonesia and Australia.

4. DISCUSSION

In this project we used a combination of balancing, model selection, and under-sampling to construct an optimized model for AIV in wild birds. We then applied this model to regions outside the sampling area in order to develop a map of the predicted relative occurrence of AIV in wild birds.

4.1. Random sub-sampling and model selection

To balance majority and minority classes in the training data and optimize performance, we employed balancing, model selection, and a combination of the two methods. We found that RSS is a simple, effective algorithm for model training and concluded that MIR alone performed poorly, and in combination with RSS, did not reliably increase model performance when classifying a rare minority class. The combination of MIR+RSS edged out RSS in the databases with 10-15% AIV-positive prevalence (on which all models performed similarly); however, MIR+RSS performed poorly on databases with 4-5% prevalence. Significant differences emerged due to exceedingly small standard deviations rather than useful differences in performance. In cases where signal was very poor, i.e. in the A3IR and ALL databases where AIV-positivity was 4-5%, RSS vastly outperformed the Default, MIR, and MIR+RSS. MIR appears to be detrimental to RSS rather than providing the synergistic boost in accuracy that we had

expected. MIR was not a prominent performance-booster, nor did it reduce OOB error rate. Perhaps this finding is not a surprise when one considers the relative immunity of Random Forests to extraneous noise in the form of non-contributing predictor variables. It should be noted that Random Forests models produced in Salford outperformed R in all cases, although the specific reasons are unclear. Salford weighting algorithms may be superior to balancing techniques implemented through randomForest in R. Interestingly, Salford produced smaller trees with fewer terminal nodes than did R, which seems to indicate that it employs a splitting method that reaches purity faster than R.

4.2. Database comparisons

The comparison of the A3IR and CIWBI databases and their combinations underscores the disadvantages of sampling across a very limited range of conditions and geographic distance. The A3IR database alone makes for such a poor training set that the resulting model performs no better than random guessing. While the CIWBI dataset by itself reaches the 0.7 AUC threshold for acceptable model accuracy, the addition of A3IR data (i.e. the ALL database) pulls the AUC values below this threshold (AUC = 0.592). It is by removing redundant points (i.e. the UNIQUE database) that the signal becomes clear (AUC = 0.766). We acknowledge that these findings are based on the data from a single project; however, it seems likely that other highly imbalanced datasets from surveillance projects that focus all their collection efforts in a very narrow area would produce similar results.

Normally, one disadvantage of under-sampling is that the removal of majority class instances involves the removal of data. Exclusion of data may withhold important

aspects from training and weaken the ability of the model to classify the majority class (He and Garcia, 2009). In the case of the UNIQUE database we did not find the removal of data to have a negative effect on the signal. Because the data that were removed were duplicates, no latitude/longitude combinations were eliminated from the model. As all the predictor variables were determined by georeferenced location, no data were lost. By combining the A3IR and CIWBI databases in this manner, we were able to construct an acceptable predictive model for AIV in wild birds. While this method of down-sampling was appopriate for this particular model, however, it may not be appropriate or useful if one were to construct a model that included bird species, biometrics, or viral strain as one would expect differences in these predictors within samples taken at the same location.

It is interesting to note that although the A3IR and CIWBI datasets share no samples in common, models constructed from these datasets performed relatively well against each other. The A3IR/RSS model scored a higher AUC on the CIWBI dataset than it did on its own data. The CIWBI/Mir + RSS model scored a slightly higher AUC on the A3IR dataset than the native model. While the reasons for these findings are unclear at this time, it emphasizes the value of cross-model comparison. Increased data-sharing would increase the pool of datasets available for this type of model evaluation.

4.3. Predictive map

Our map predicts regions of high relative occurrence primarily located in high-latitude regions with a continental climate. Previous studies have demonstrated the increased predicted risk of AIV in northern regions using high-latitude data (Chapter 2, this volume) as well as global data (Chapter 3, this volume).

4.4. Important variables

The importance of landcover type was consistent with the findings of a project that mapped AIV in the continental United States (Fuller et al., 2010) and found that the amount of harvested cropland per county was an important predictor. Croplands attract waterfowl, and during migration, agricultural crops make up a large percentage of their diet (Miller et al., 2000). A GARP model predicted presence of HPAI H5N1 in woodlands and absence in coastal mangroves and freshwater swamps (Williams et al., 2008). Temperature seasonality appeared as an important factor in two HPAI H5N1 models constructed with GARP, but differed in the nature of this variable. Adhikari et al. (2009) found that a very low seasonal variation of 5° C was correlated with HPAI H5N1 outbreak in domestic chickens in India (Adhikari et al., 2009). Williams' (2008) model predicted high seasonal variation and dry climate. Where temperature is concerned, our best model is consistent with others in showing that colder temperatures are correlated with increased in the predicted occurrence of AIV. Another large-scale predictive map of AIV in wild birds of the continental United States identified cold temperatures (i.e. early freeze date) as an important predictor for AIV risk in wild birds (Fuller et al., 2010).

One major difference between the map we present in this project and our two previous, wide-scale models of AIV in wild birds (Chapters 2 and 3, this volume) is that in this current map, temperate and tropical regions show areas of increased predicted occurrence, whereas the two prior maps had a wide equatorial band of low predicted occurrence. In particular, Indonesia displays a mottled pattern of moderate to high risk, which is consistent with reported H5N1 outbreaks (Kandun et al., 2006) and a declaration

by the FAO that Indonesia is the country most affected by AIV (Northoff, 2008). As Indonesia is clearly not a region characterized by bitterly cold winters and cool summers, it may share aspects with the areas that caused the secondary peaks seen in the partial dependence plots, which were typically in the very warm range. These jagged, secondary peaks are unique to this study and were not seen in previous models (Chapters 2 and 3, this volume). Whether these secondary peaks and mottled patterns are weak signal or noise requires more study and highlights some gaps that could be filled with targeted surveillance effort, particularly sampling of migratory waterfowl in low-latitude regions.

4.5 Conclusions

Yearly sampling may serve the purpose of observing trends over time in bird species or viral strain fluctuations at a single location. However, it appears that concentrating collection efforts and amassing data points in a very limited area not only swamps any signal, but also skews the range of the predictor variables such that it is almost impossible to construct an accurate prediction model and extrapolate findings beyond the sampling location. If surveillance efforts continue to be carried out in the traditional manner in traditional locations without regard for basic principles of sampling theory, one can expect diminishing returns in the heuristic value of the data collected, while the effort and cost of field and laboratory work remain the same. Data-mining and machine-learning applications are then left to glean new results from what is essentially the same data. Adaptive Management principles for conservation are highly applicable in this case (Nichols and Williams, 2006). Collection in the field and laboratory testing would occur in cooperation with machine-learning and data-mining projects; this arrangement would

be mutually beneficial. Existing surveillance data are used as a baseline to identify gaps in knowledge that pertain to specific, quantifiable research questions. Data-mining and machine-learning results guide hypothesis-building, field studies, and laboratory work. The results would strengthen existing models and provide material for new models; then, new and more accurate models would be the basis for new hypotheses, field studies and laboratory work (Fig. 4.9). Furthermore, modeling is a convenient "quality control" step that will identify gaps in surveillance effort. With each iterative cycle, the information produced becomes more and more useful and accurate. While this process requires more planning, experimental design, hypothesis development, and general scientific effort, the results obtained from targeted collection will provide specific information applicable to decision making, risk assessment, and disease management. The complicated ecology of AIV requires this cycle of targeted questioning, monitoring, and modeling to maximize limited resources and usefulness of results.

ACKNOWLEDGEMENTS

We thank Dr. Jane Parmley and the Canadian Cooperative Wildlife Health Centre for information about the CIWBI database. We thank Mohammed Khalilia for helpful discussions on implementing RSS in R, and Michael A. Lindgren for expertise and valuable input in data analysis. We acknowledge and thank Dr. Eric Bortz, Florian Aldehoff, and all CEIRS collaborators for their contribution to AIV surveillance.

TABLES

Table 4.1. Selected examples of the prevalence of birds testing positive for avian influenza virus (AIV) from wild bird surveillance projects.

Bird type	Location	Season	AIV Prevalence	Reference
Juvenile wild ducks	Alberta, Canada	August	18-60%	(Hinshaw et al., 1980)
Northern shovelers	Breeding areas	August/September	32%	(Hill et al., 2012)
Live wild ducks	Canada	July-December	30%	(Parmley et al., 2009)
Wild ducks	Alberta, Canada	July - September	22%	(Krauss et al., 2004)
Wild ducks	Netherlands, Sweden	Peak prevalence in September/October	7%	(Munster et al., 2007)
Northern shovelers	California	Winter	5%	(Hill et al., 2012)
Shorebirds	Delmarva Peninsula, USA	July – September	0-5%	(Krauss et al., 2004)
Migratory waterfowl	China	October – March	>1%	(Chen et al., 2006)
Passerines	Continental U.S.	Fall/Winter	>1%	(Fuller et al., 2010)
Charadriiformes	Netherlands, Sweden	Peak prevalence in September/October	>1%	(Munster et al., 2007)

Table 4.2. Predictor variables used by the Random Forests to create a prediction map for AIV in wild birds. Variables are listed in order of their Variable Importance Score (VIS) or normalized relative contribution to model accuracy. These VIS reflect the analysis by the MIR+RSS method, so predictors with an "NA" score for VIS were excluded from final analysis.

Predictor variable	Description	VIS	Source
Mean temperature, April	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	100.0	(Hijmans et al., 2005)
Landcover	20 categories	99.3	(Bartholomé and Belward, 2005)
Temperature seasonality	Standard deviation x 100	95.0	(Hijmans et al., 2005)
Distance from river, lake, wetland	Calculated from combined large and small lake polygons, and lakes and wetlands grid	92.7	(Lehner and Döll, 2004)
Slope	In degrees	90.3	Calculated from elevation
Temperature annual range	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	84.7	(Hijmans et al., 2005)
Human population density	Population density for 2010, 2.5' resolution, persons/km^2	83.6	(Center for International Earth Science Information Network (CIESIN) et al., 2005)
Human impact index	Summative index of human disturbance (0-72)	82.5	(Sanderson et al., 2002)
Mean temperature, February	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	81.5	(Hijmans et al., 2005)
Elevation	In m; 30 arc-seconds, 1 km spatial resolution	80.5	"
Mean temperature, November	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	77.5	(Hijmans et al., 2005)
Mean temperature, June	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	76.1	"
Precipitation of warmest quarter	In mm; 30 arc-seconds, 1 km spatial resolution	75.5	"
Precipitation of wettest quarter	In mm; 30 arc-seconds, 1 km spatial resolution	74.6	"
Mean temperature, December	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	73.7	"
Mean temperature, July	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	73.6	"
Mean temperature of coldest quarter	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	73.2	"

Table 4.2 continued

Predictor variable	Description	VIS	Source
Mean temperature, May	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	71.9	(Hijmans et al., 2005)
Mean diurnal range	Mean of monthly(max temp-min temp)	69.5	"
Mean temperature, March	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	67.8	"
Mean temperature, October	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	66.5	"
Min temperature of coldest month	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	65.4	"
Mean temperature, October	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	65.2	"
Max temperature of coldest month	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	65.2	"
Mean temperature of driest quarter	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	62.3	"
Distance from coast	In m; calculated from coastline	58.8	"
Precipitation of wettest month	In mm; 30 arc-seconds, 1 km spatial resolution	58.4	"
Mean temperature, September	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	56.0	"
Mean temperature, August	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	55.9	"
Precipitation seasonality	standard deviation of weekly estimates expressed as a percentage of the mean	55.6	"
Annual precipitation	In mm; 30 arc-seconds, 1 km spatial resolution	52.8	"
Annual mean temperature	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	50.1	"
Precipitation of coldest quarter	In mm; 30 arc-seconds, 1 km spatial resolution	45.5	"
Precipitation of driest month	In mm; 30 arc-seconds, 1 km spatial resolution	43.4	"
Precipitation of driest quarter	In mm; 30 arc-seconds, 1 km spatial resolution	41.6	"
Isothermality	Mean diurnal range/Annual temperature range; In ^{o}C; 30 arc-seconds, 1 km spatial resolution	NA	"
Mean temperature of wettest quarter	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	NA	"
Mean temperature of warmest quarter	In ^{o}C; 30 arc-seconds, 1 km spatial resolution	NA	"

Table 4.3. Descriptive summary table for databases.

Database	Contributor	# Samples	AIV Prevalence
A3IR	Alaska Asia Avian Influenza Research Group	40,741	4%
CIWBI	Canada's Inter-agency Wild Bird Influenza Survey	4,298	15%
ALL	A3IR and CIWBI combined	45,039	5%
UNIQUE	Redundant AIV-negative lat/lon combinations removed	1,930	10%

Table 4.4. Summary table for experimental methods.

Method	Description	Details
Default	Default	In R using default settings/values. Number of trees = 500, number of predictors at each node = sqrt(total number of predictors).
MIR	Model Improvement Ratio	In R. Using the variable importance scores from the intial Default run, stepwise, descending elimination of variables was performed to find the subset of predictors that produced the lowest OOB error.
RSS	Repeated Random Sub-Sampling	Random Forests is run in R with a sub-sample of data made up of all the positive points and an equal number of randomly selected negative points. One tree is constructed with this subset, and then negative points are sampled with replacement. 500 trees are constructed and their results are aggregated into a single object using the "combine" command in R.
MIR+RSS	MIR & RSS	The MIR algorithm is run in R with repeated random sub-sampling at each threshold.
Salford	Salford Predictive Miner	Random Forests analysis is performed with Salford Predictive Modeler using the following settings: Tree Type = Classification, Class Weights = BALANCED, Number of trees = 500, Number of predictors for each node = 6.

Table 4.5. Descriptive statistics for databases and models. This table presents mean area under the receiver operating characteristic curve (AUC), standard error, 95% confidence intervals based on t-distributions, mean OOB error rate, and mean treesize (number of nodes). The p-values indicate the probability that these values differ from a random model (AUC = 0.5) by chance. The highest scoring AUC values (excluding Salford results) are in red. [A]$p < 0.005$, n = 10. [B]$p < 0.05$, n=10. [C]$p<0.001$, n=10. [D]$p<0.05$, n=10. The Salford rows represent results of a single forest.

				95% confidence interval				
		AUC	**SE**	lower bound	upper bound	**p <**	**OOB**	**Treesize**
A3IR	Default	0.490	0.007	0.473	0.506	0.195	0.039	118
	MIR	0.485	0.005	0.473	0.497	0.018	0.039	39
	RSS	**0.542**[A]	0.010	0.519	0.565	0.002	0.032	118
	MIR+RSS	0.494	0.004	0.484	0.504	0.195	0.039	121
	Salford	0.645	--	--	--	--	--	16
CIWBI	Default	0.694	0.003	0.687	0.700	0.000	0.145	182
	MIR	0.687	0.004	0.678	0.696	0.000	0.145	181
	RSS	0.686[B]	0.002	0.681	0.691	0.000	0.261	182
	MIR+RSS	**0.697**[B]	0.003	0.690	0.703	0.000	0.145	182
	Salford	0.837	--	--	--	--	--	59
ALL	Default	0.592	0.002	0.587	0.597	0.000	0.049	62
	MIR	0.591	0.001	0.589	0.593	0.000	0.049	62
	RSS	**0.668**[A]	0.007	0.653	0.684	0.000	0.037	62
	MIR+RSS	0.586	0.006	0.572	0.599	0.000	0.049	61
	Salford	0.720	--	--	--	--	--	81
Unique	Default	0.766	0.001	0.764	0.768	0.000	0.102	125
	MIR	0.765[D]	0.001	0.762	0.767	0.000	0.105	130
	RSS	0.760[C,D]	0.001	0.758	0.762	0.000	0.172	125
	MIR+RSS	**0.767**[D]	0.001	0.766	0.768	0.000	0.102	126
	Salford	0.793	--	--	--	--	--	83

FIGURES

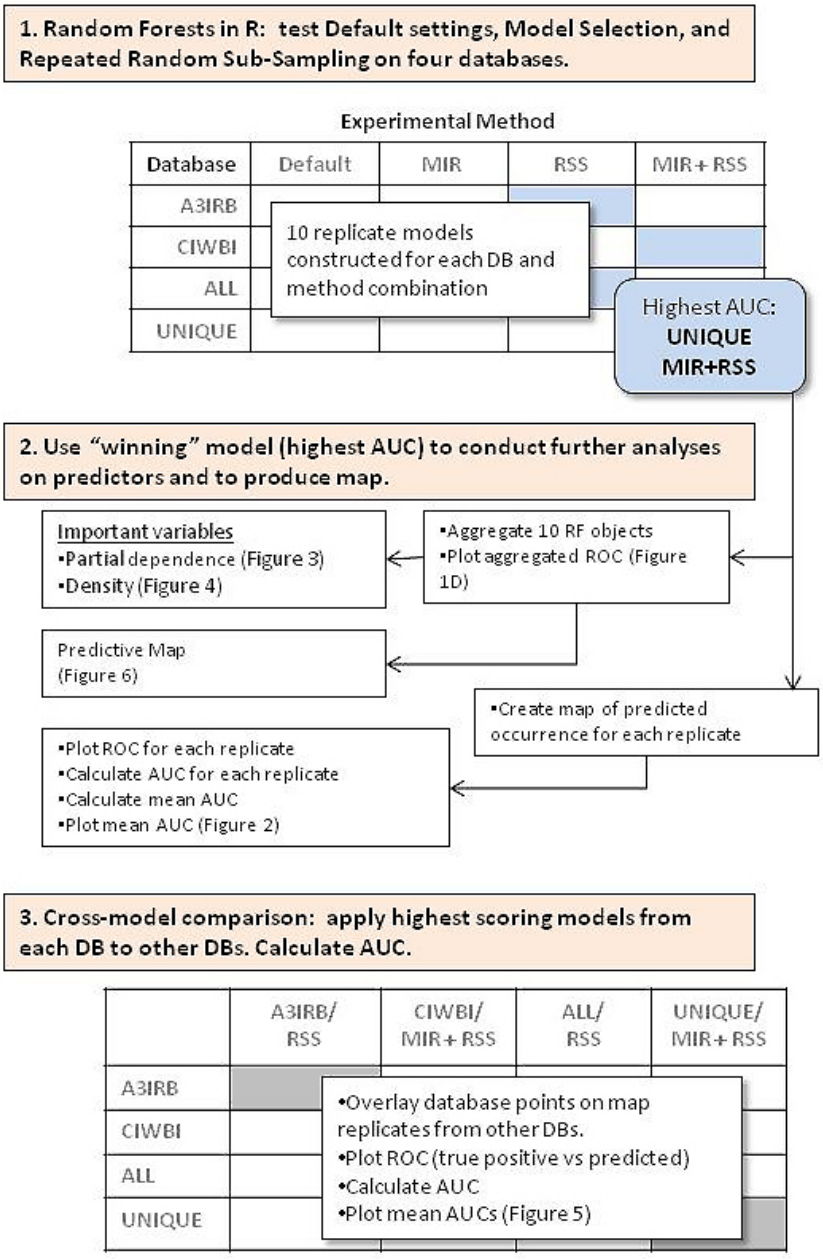

Figure 4.1. Research design. This diagram follows the path of the UNIQUE/MIR+RSS model. Altogether, 16 models were constructed and the best model for each database underwent the same steps, except for the production of the predictive map and the analyses of the important variables. Salford results could not be analyzed in a parallel manner so they are excluded from this diagram.

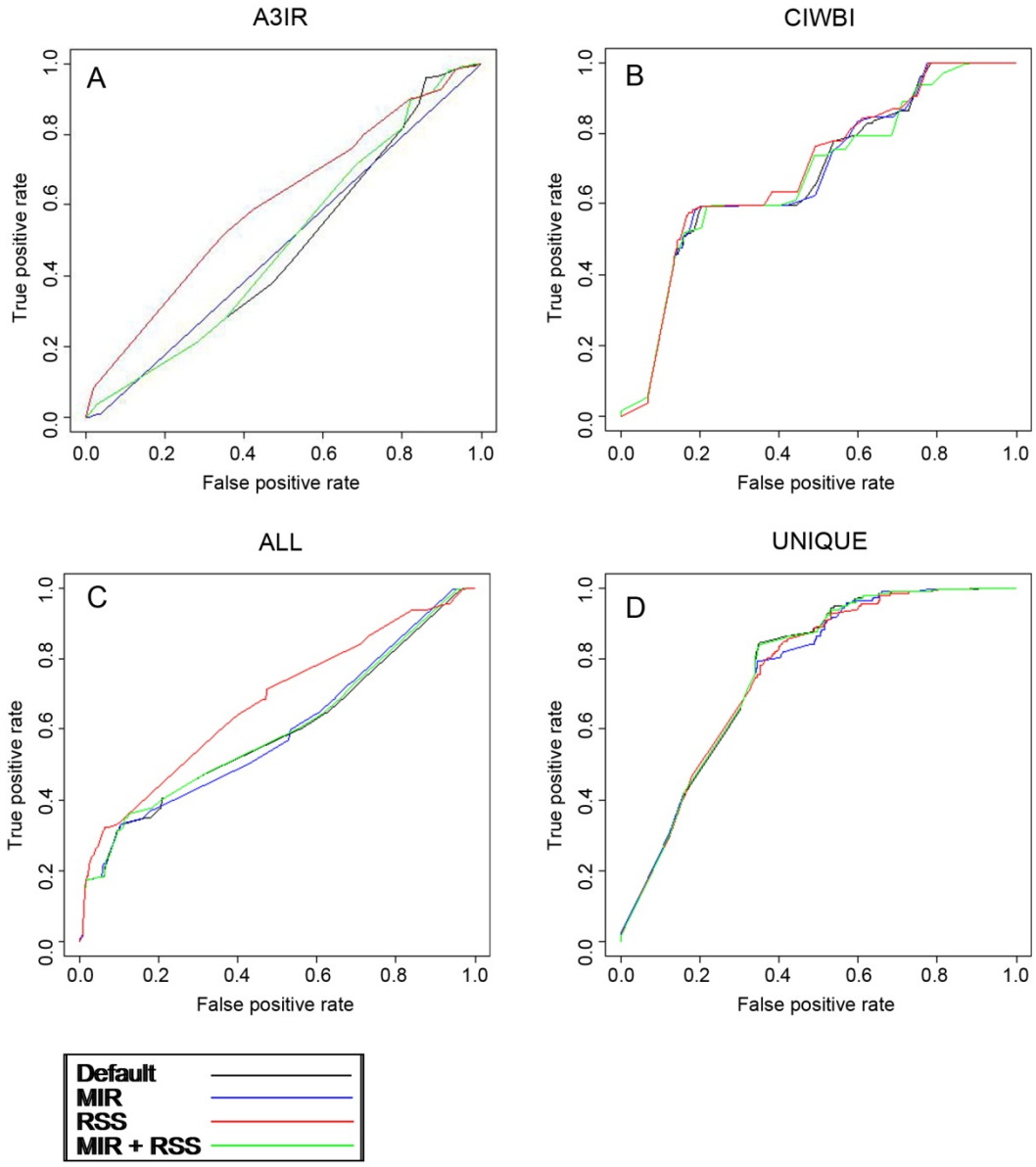

Figure 4.2. Receiver Operating Characteristic (ROC) curves for experimental methods. The ROC from the Default (black), model improvement ratio (MIR; blue), repeated random sub-sampling (RSS; red), and MIR + RSS (green) methods are grouped by the datasets they were applied to. Datasets include the Alaska Asia Avian Influenza 2005-2010 (A3IR; A), Canada's Inter-agency Wild Bird Influenza survey (CIWBI; B), the two databases combined (ALL; C), and the unique latitude/longitude values from ALL (UNIQUE; D). ROC is generated by plotting true positives against false positives. ROC curves that are no different from a diagonal line (i.e. A, C) indicate that the models predicted no better than random assignment.

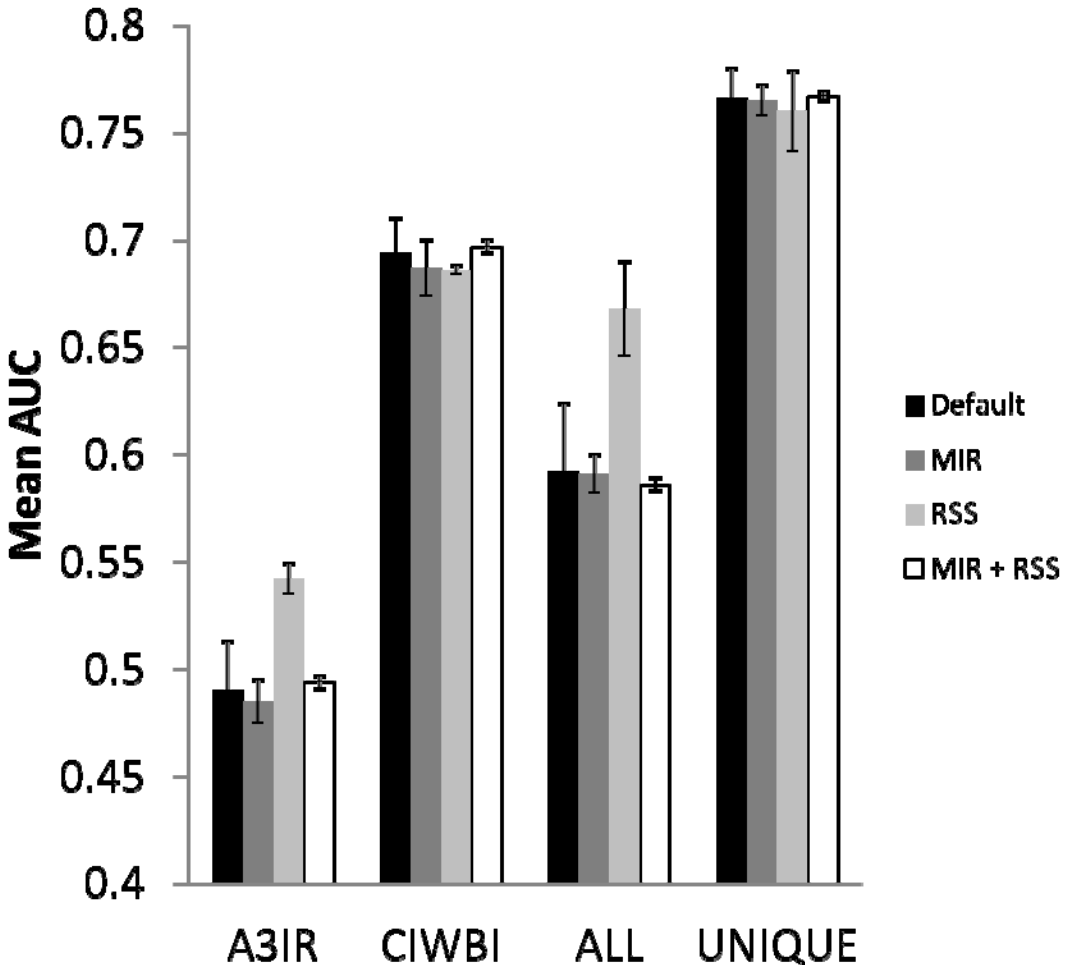

Figure 4.3. Mean area under the receiver operating characteristic curves (AUC) of the four different experimental methods that generated them: Default (black), model improvement ratio (MIR; dark gray), repeated random sub-sampling (RSS; light gray), and MIR + RSS (white). Methods are grouped by the databases to which they were applied: Alaska Asia Avian Influenza 2005-2010 (A3IR), Canada's Inter-agency Wild Bird Influenza survey (CIWBI), the two databases combined (ALL), and the unique latitude/longitude values from ALL (UNIQUE). Each bar represents mean AUC (n=10) $\pm$ standard error. For specific values and significant differences are presented in Table 5. Salford results could not be analyzed in a parallel manner so they are excluded from this diagram. A model generating a mean AUC at or above 0.7 is considered to have acceptable predictive power.

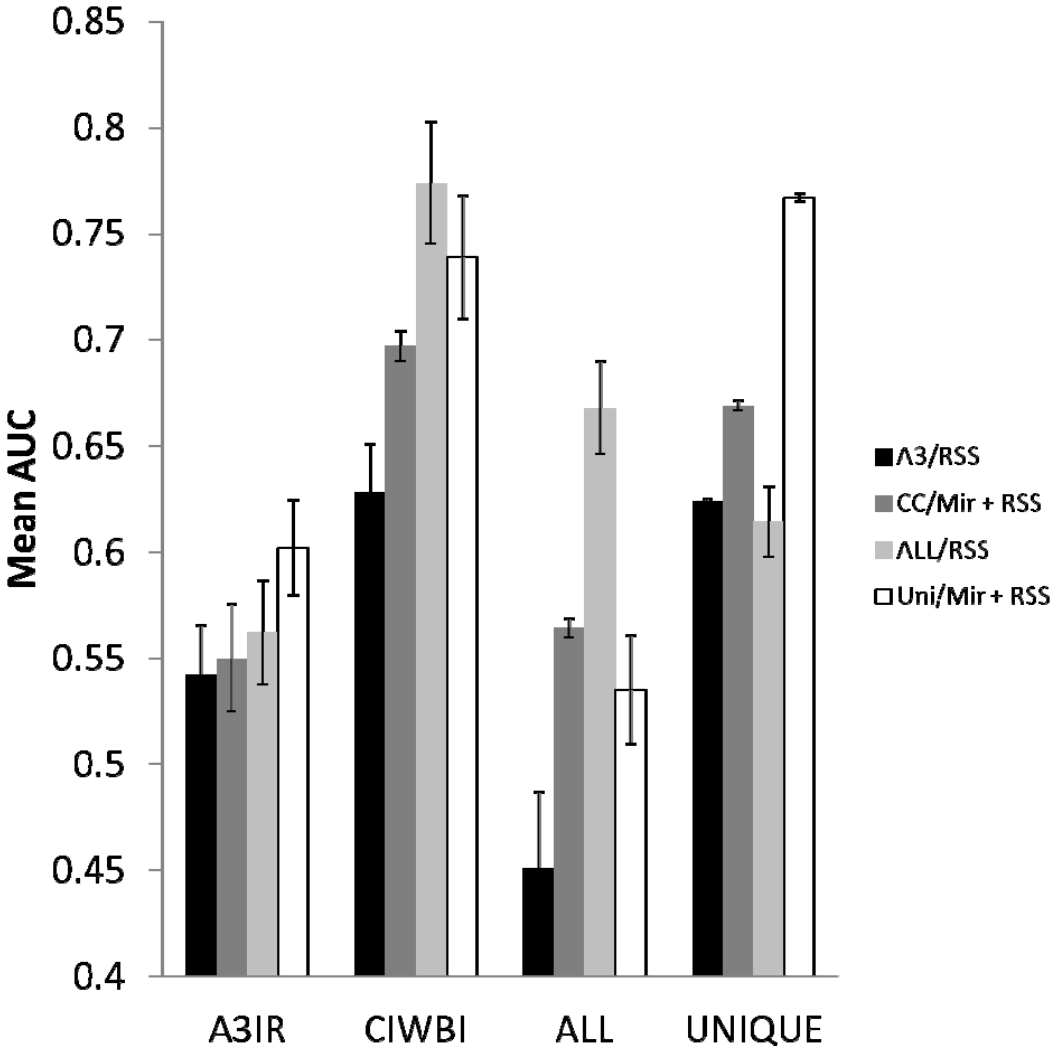

Figure 4.4. Cross-model comparison results. Mean area under the receiver operating characteristic curve (AUC) ± standard error representing the predictive success of the four top-scoring models: repeated random sub-sampling (RSS) applied to the Alaska Asia Avian Influenza Research (A3IR) database (A3/RSS); model improvement ratio (MIR) combined with RSS applied to the Canada's Inter-agency Wild Bird Influenza survey (CIWBI) database (CC/Mir + RSS); RSS applied to the combined A3IR and CIWBI databases (ALL; ALL/RSS); and MIR and RSS applied to a database composed of the unique latitude/longitude coordinates from ALL (UNIQUE; Uni/Mir + RSS). Models are grouped by the databases they were applied to (A3IR, CIWBI, ALL, UNIQUE). A model generating a mean AUC at or above 0.7 is considered to have acceptable predictive power.

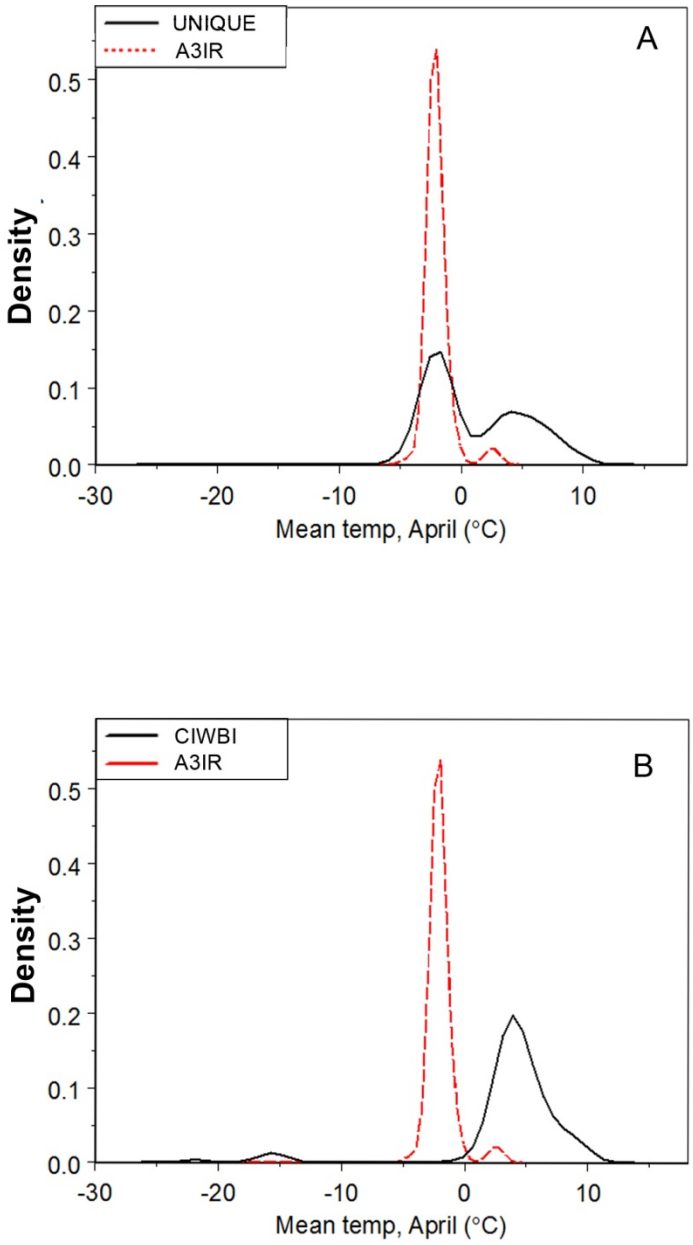

Figure 4.5. Density plots for the mean temperature in April. These two plots contrast the very narrow sampling range found in the Alaska Asia Avian Influenza Research (A3IR) database with those in the UNIQUE databases (unique latitude/longitude coordinates from the combined A3IR and CIWBI dataset; A) and Canada's Inter-agency Wild Bird Influenza survey (CIWBI; B).

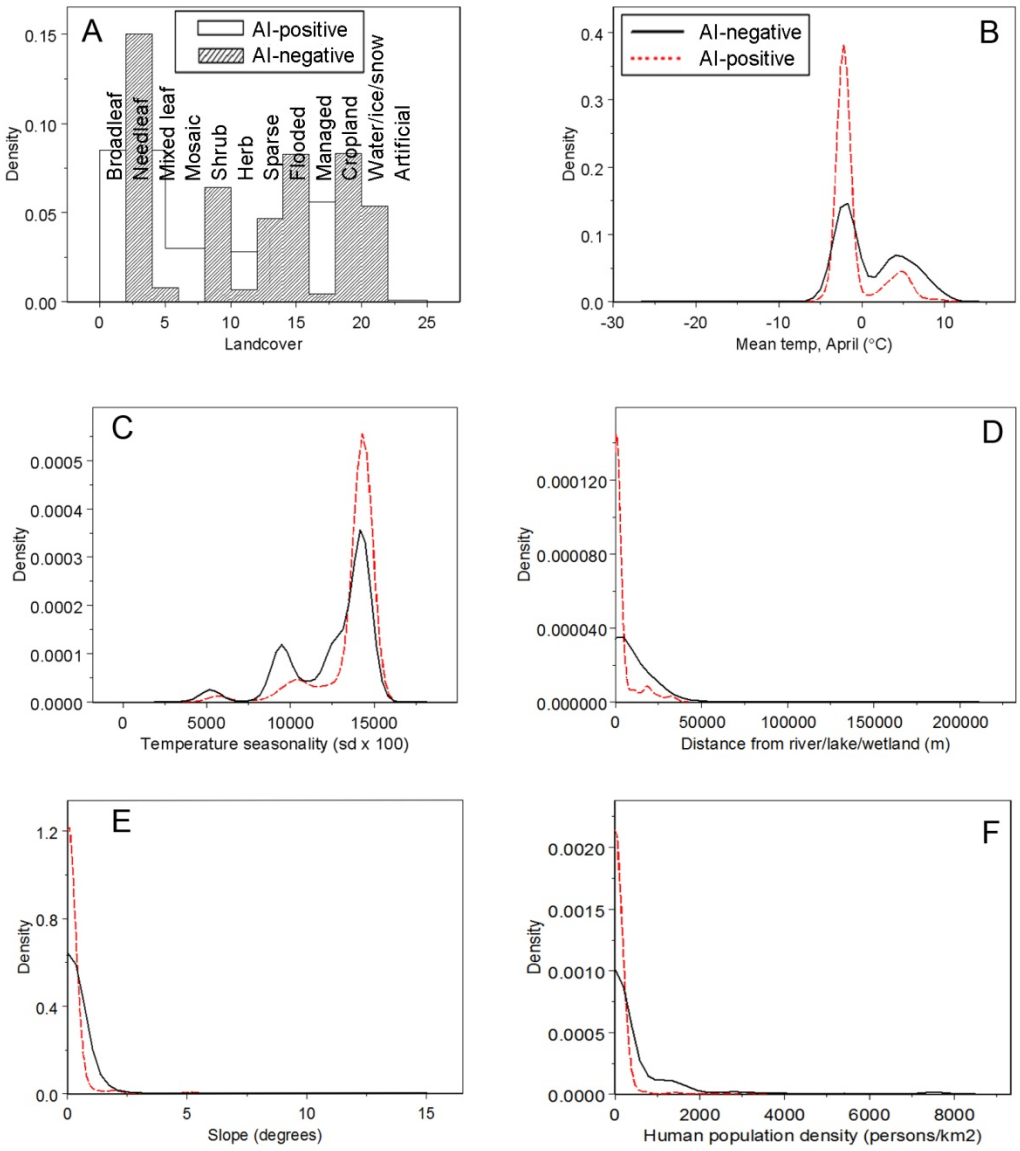

Figure 4.6. Density plots for important variables. Landcover (A), Mean temperature in April (°C ,B), Temperature seasonality (°C x 100; C), Distance from river/lake/wetland (m; D), Slope (degrees; E), and Human population density (persons/km^2; F) were the highest contributors to model accuracy as calculated by Random Forests. Density lines (or bars, as in A) for cases that tested positive for avian influenza virus (AIV) are represented by dashed lines, AIV-negative by solid lines. Peaks in AIV-positive cases that exceed AIV-negative cases indicate a range of values for this predictor variable that are correlated with AIV-positivity.

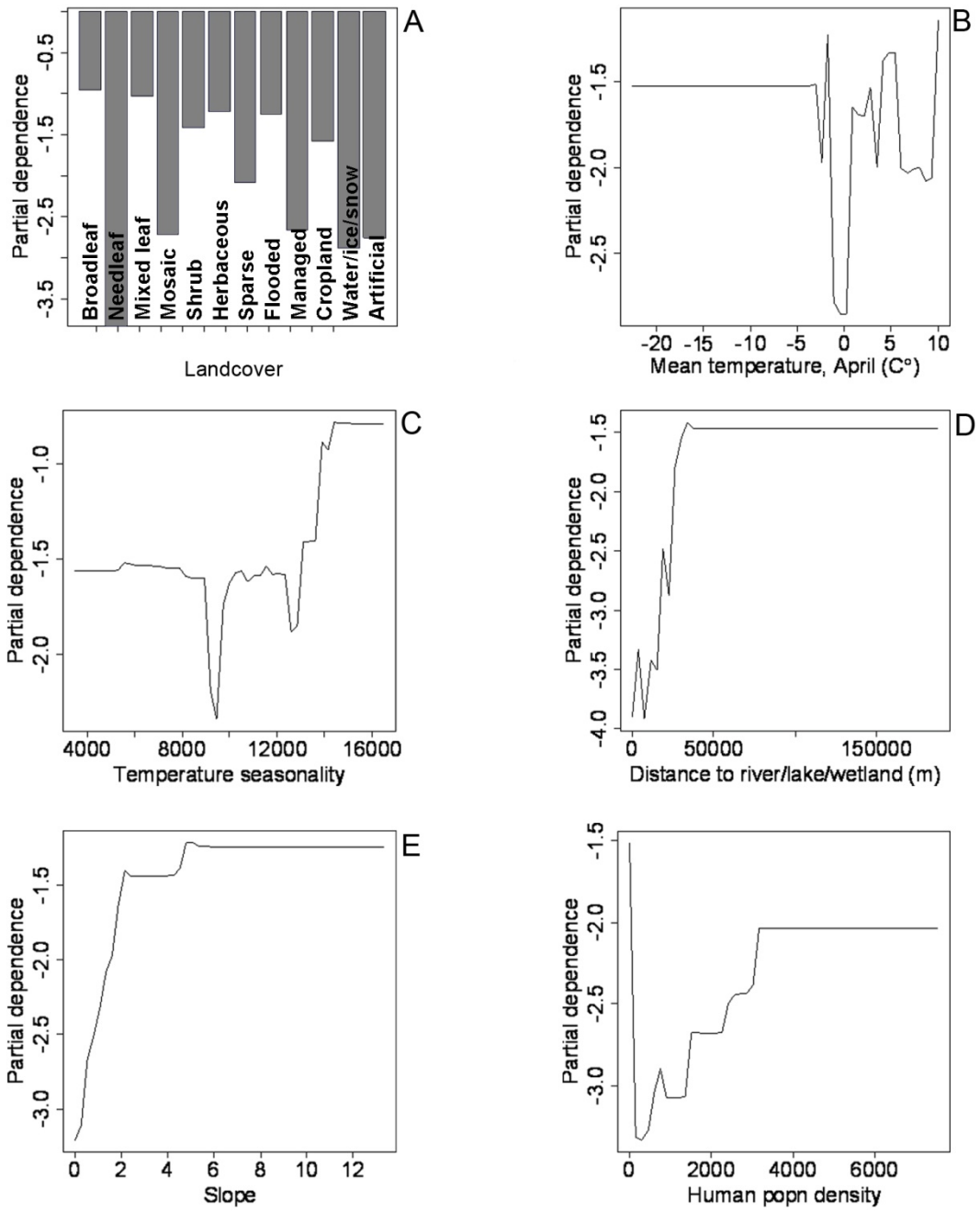

Figure 4.7. Partial dependence plots for important predictor variables. Landcover (A), Mean temperature in April ((°C; B), Temperature seasonality (°C x 100; C), Distance from river/lake/wetland (m; D), Slope (degrees; E), and Human population density (persons/km^2; F) were the highest contributors to model accuracy as calculated by Random Forests. Plots show the partial dependence of avian influenza virus (AIV)-positivity on each predictor variable over the averaged effects of all other predictors.

Partial dependence is best interpreted as an index of the importance of values within a variable's range and is best understood by examining general patterns in relation to the values of the predictor variable rather than the specific values of partial dependence. For example, in A, which is displayed as bars because Landcover is a categorical variable, AIV-positivity is most highly dependent on needleleaf landcover, in B the dependence on the mean temperature in April is moderate at ranges below -5 °C, variable between 0 and 10 °C where partial dependence spikes. The variability may be due to interactions with other predictors. In F, partial dependence on human population density is highest at 0 persons/km^2, then drops and gradually increases to a moderate plateau around 4000 person/km^2.

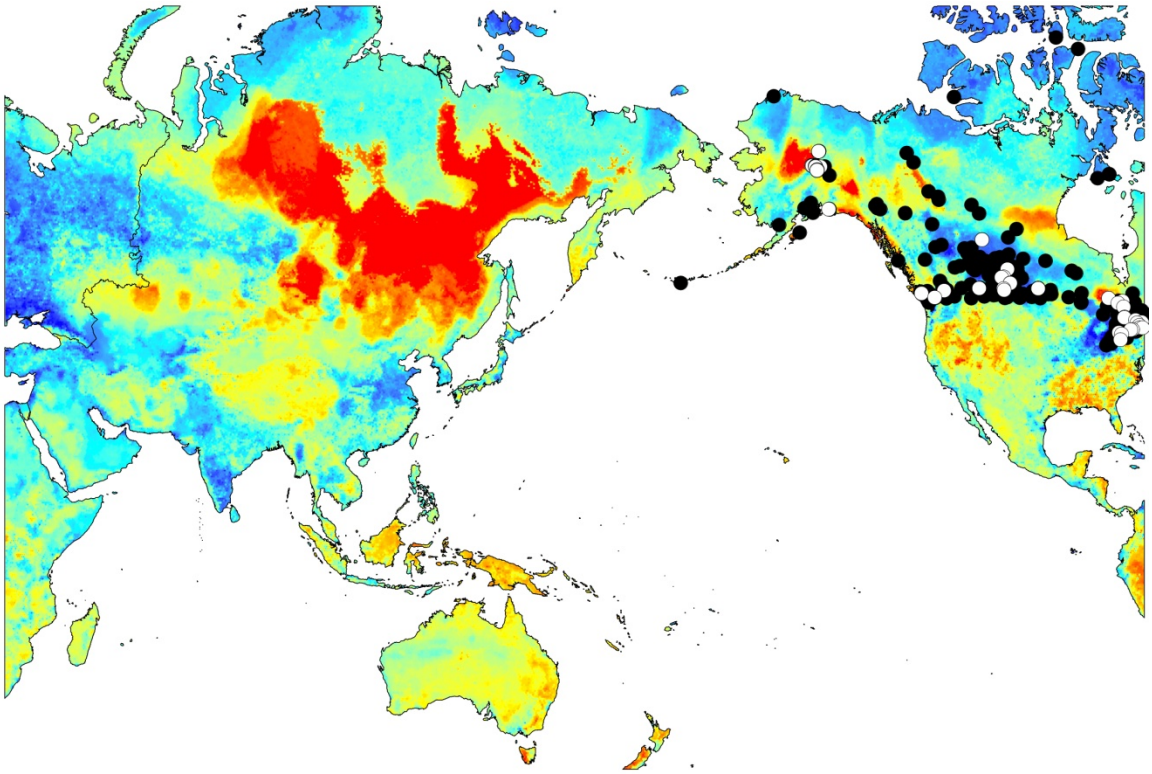

Figure 4.8. Map of predicted relative occurence index of avian influenza virus (AIV) in wild birds and surveillance locations around the Pacific Rim study area. Map was generated by the model improvement ratio/repeated random sub-sampling method applied to a database of points containing only unique latitude/longitude coordinates. The map colors indicate areas of high (red) predicted occurrence of AIV to low predicted occurrence (blue). Collection locations are indicated by dots: white dots represent locations where at least one AIV-positive sample was collected; black dots are locations where only AIV-negative samples were collected.

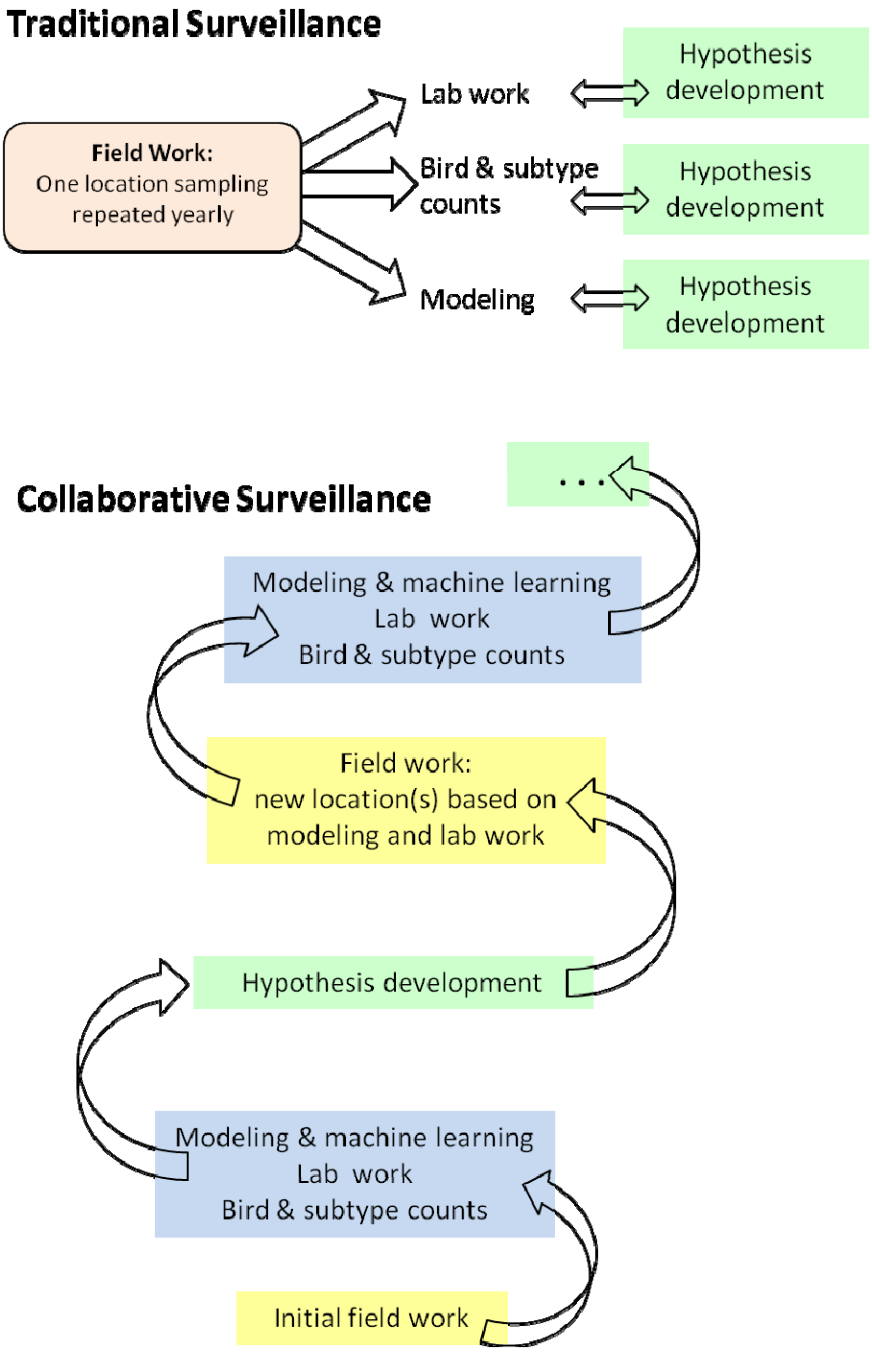

Figure 4.9. A conceptual diagram illustrating differences between traditional and collaborative surveillance methods and their interaction with laboratory and machine-learning work.

LITERATURE CITED

Abolnik, C., Gerdes, G., Sinclair, M., Ganzevoort, B., Kitching, J., Burger, C., 2010. Phylogenetic analysis of influenza A viruses (H6N8, H1N8, H4N2, H9N2, H10N7) isolated from wild birds, ducks, and ostriches in South Africa from 2007 to 2009. Avian Dis. 54, 313-322.

Adhikari, D., Chettri, A., Barik, S.K., 2009. Modelling the ecology and distribution of highly pathogenic avian influenza (H5N1) in the Indian subcontinent. Curr. Sci. 97, 72-78.

Bartholomé, E., Belward, A.S., 2005. GLC2000: a new approach to global land cover mapping from Earth observation data. Int. J. Remote Sens. 26, 1959-1977.

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5-32.

Cecchi, G., Ilemobade, A., Brun, Y.L., Hogerwerf, L., Slingenbergh, J., 2008. Agro-ecological features of the introduction and spread of the highly pathogenic avian influenza (HPAI) H5N1 in northern Nigeria. Geospat. Health 3, 7-16.

Center for International Earth Science Information Network (CIESIN) & Columbia University; and Centro Internacional de Agricultura Tropical (CIAT). Gridded Population of the World Version 3 (GPWv3): Population Density Grids, http://sedac.ciesin.columbia.edu/gpw.

Chen, C., Liaw, A., Breiman, L., 2004. Using Random Forest to learn imbalanced data, Technical Report 666. Department of Statistics, University of California, Berkeley, CA, pp. 1-12.

Chen, H., Smith, G.J.D., Li, K.S., Wang, J., Fan, X.H., Rayner, J.M., Vijaykrishna, D., Webster, R.G., 2006. Establishment of multiple sublineages of H5N1 influenza virus in Asia: Implications for pandemic control. Proc. Natl. Acad. Sci. U.S.A. 103, 2845-2850.

Cilloni, F., Toffan, A., Giannecchini, S., Clausi, V., Azzi, A., Capua, I., Terregino, C., 2010. Increased pathogenicity and shedding in chickens of a wild bird-origin low pathogenicity avian influenza virus of the H7N3 subtype following multiple in vivo passages in quail and turkey. Avian Dis. 54, 555-557.

Craig, D., Huettmann, F., 2008. Chapter IV, Using "blackbox" algorithms such as TreeNet and RandomForest for data-mining and for finding meaningful patterns, relationships and outliers in complex ecological data, In: Wang, H.-f. (Ed.), Intelligent Data Analysis: Developing New Methodologies through Pattern Discovery and Recovery. IGI Global, Hershey, pp. 65-84.

Cutler, D.R., Edwards, T.C., Jr., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random Forests for classification in ecology. Ecology 88, 2783-2792.

Davison, S., Eckroade, R.J., Ziegler, A.F., 2003. A Review of the 1996–98 Nonpathogenic H7N2 avian influenza outbreak in Pennsylvania. Avian Dis. 47, 823-827.

Evans, J.S., Cushman, S.A., 2009. Gradient modeling of conifer species using random forests. Landscape Ecol. 24, 673-683.

Fawcett, T., 2004. ROC graphs: notes and practical considerations for researchers, Technical Report HPL-2003-4. HP Labs, Palo Alto, CA, pp. 1-38.

Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 24, 38-49.

Freeman, E.A., Moisen, G.G., Frescino, T.S., 2012. Evaluating effectiveness of down-sampling for stratified designs and unbalanced prevalence in Random Forest models of tree species distributions in Nevada. Ecol. Model. 233, 1-10.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29, 1189-1232.

Fuller, T.L., Saatchi, S.S., Curd, E.E., Toffelmeier, E., Thomassen, H.A., Buermann, W., Smith, T.B., 2010. Mapping the risk of avian influenza in wild birds in the U.S. BMC Infect. Dis. 10, 187.

Genuer, R., Poggi, J.-M., Tuleau-Malot, C., 2010. Variable selection using Random Forests. Pattern Recog. Lett. 31, 2225-2236.

Gilbert, M., Chaitaweesub, P., Parakamawongsa, T., Premashthira, S., Tiensin, T., Kalpravidh, W., Wagner, H., Slingenbergh, J., 2006. Free-grazing ducks and highly pathogenic avian influenza, Thailand. Emerg. Infect. Dis. 12, 227-234.

Gilbert, M., Xiao, X., Pfeiffer, D.U., Spprecht, M., Boles, S., Czarnecki, C., 2008. Mapping H5N1 highly pathogenic avian influenza risk in Southeast Asia. Proc. Natl. Acad. Sci. U.S.A. 105, 4769-4774.

Globig, A., Staubach, C., Beer, M., Koeppen, U., Fielder, W., Nieburg, M., Harder, T.C., 2009. Epidemiological and ornithological aspects of outbreaks of highly pathogenic avian influenza virus H5N1 of Asian lineage in wild birds in Germany, 2006 and 2007. Transbound. Emerg. Dis. 56, 57-72.

Goyal, S.M., Jindal, N., Chander, Y., Ramakrishnan, M.A., Redig, P.T., Sreevatsan, S., 2010. Isolation of mixed subtypes of influenza A virus from a bald eagle (*Haliaeetus leucocephalus*). Virol. J. 7, 174.

Hall, J.S., Ip, H.S., Franson, J.C., Meteyer, C., Nashold, S., TeSlaa, J.L., 2009. Experimental infection of a North American raptor, American kestrel (*Falco sparverius*), with highly pathogenic avian influenza virus (H5N1). PLoS ONE 4, e7555.

He, H., Garcia, E.A., 2009. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21, 1263-1284.

Hegel, T.M., Cushman, S.A., Evans, J., Huettmann, F., 2010. Current state of the art for statistical modelling of species distributions, In: Cushman, S.A., Huettmann, F. (Eds.), Spatial Complexity, Informatics, and Wildlife Conservation. Springer, Tokyo, pp. 273-311.

Herrick, K.A., Huettmann, F., Lindgren, M.A., 2013. A global model of avian influenza prediction in wild birds: the importance of northern regions. Vet. Res. submitted.

Herrick, K.A., Huettmann, F., Runstadler, J., Chernetsov, N., Antonov, A., Valchuk, O., Gerasimov, Y., Matsyna, E., Matsyna, A., Markovets, M., Druzyaka, A., Saito, K., 2010. Predictive RISK modeling of avian influenza in the Pacific Rim and beyond, In: Kremers, H., Susini, A. (Eds.), Risk Models and Applications, 2010. CODATA Germany: Lecture Notes in Information Sciences, Berlin, pp. 135-148.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25, 1965-1978.

Hill, N.J., Takekawa, J.Y., Cardona, C.J., Meixell, B.W., Ackerman, J.T., Runstadler, J.A., Boyce, W.M., 2012. Cross-seasonal patterns of avian influenza virus in breeding and wintering migratory birds: a flyway perspective. Vector Borne Zoonotic Dis. 12, 243-253.

Hinshaw, V.S., Webster, R.G., Turner, B., 1980. The perpetuation of orthomyxoviruses and paramyxoviruses in Canadian waterfowl. Can. J. Microbiol. 26, 622-629.

Hogerwerf, L., Walker, R.G., Ottaviani, D., Slingenbergh, J., Prosser, D., Bergmann, L., Gilbert, M., 2010. Persistence of highly pathogenic avian influenza H5N1 virus defined by agro-ecological niche. EcoHealth 7, 213-225.

Hosmer, D.W., Lemeshow, S., 2000. Applied Logistic Regression. Wiley, New York.

Ito, T., Goto, H., Yamamoto, E., Tanaka, H., Takeuchi, M., Kuwayama, M., Kawaoka, Y., Otsuki, K., 2001. Generation of a highly pathogenic avian influenza A virus from an avirulent field isolate by passaging in chickens. J. Virol. 75, 4439-4443.

Jourdain, E., Gauthier-Clerc, M., Sabatier, P., 2007. Ecoregional dominance in spatial distribution of avian influenza (H5N1) outbreaks [response]. Emerg. Infect. Dis. 13, 1270-1271.

Kaleta, E., Pena, K., Yilmaz, A., Redmann, T., Hofheinz, S., 2007. Avian influenza A viruses in birds of the order Psittaciformes: reports on virus isolations, transmission experiments and vaccinations and initial studies on innocuity and efficacy of oseltamivir in ovo. Dtsch. Tierarztl. Wochenschr. 114, 260-267.

Kandun, I.N., Wibisono, H., Sedyaningsih, E.R., Yusharmen, D.P.H., Hadisoedarsuno, W., Purba, W., 2006. Three Indonesian clusters of H5N1 virus infection in 2005. N. Engl. J. Med. 355, 2186-2194.

Khalilia, M., Chakraborty, S., Popescu, M., 2011. Predicting disease risks from highly imbalanced data using random forest. BMC Med. Inf. Decis. Making 11, 51-63.

Kilpatrick, A.M., Chmura, A.A., Gibbons, D.W., Fleischer, R.C., Marra, P.P., Daszak, P., 2006. Predicting the global spread of H5N1 avian influenza. Proc. Natl. Acad. Sci. U.S.A. 103, 19368-19373.

Krauss, S., Obert, C.A., Franks, J., Walker, D., Jones, K., Webster, R.G., 2007. Influenza in migratory birds and evidence of limited intercontinental virus exchange. PLoS Path. 3, e167.

Krauss, S., Walker, D., Pryor, P., Niles, L., Chenghong, L., Hinshaw, V.S., Webster, R.G., 2004. Influenza A viruses of migrating wild aquatic birds in North America. Vector Borne Zoonotic Dis. 4, 177-189.

Lee, C.-W., Swayne, D.E., Linares, J.A., Senne, D.A., Suarez, D.L., 2005. H5N2 avian influenza outbreak in Texas in 2004: the first highly pathogenic strain in the United States in 20 years? J. Virol. 79, 11412-11421.

Lehner, B., Döll, P., 2004. Development and validation of a global database of lakes, reservoirs and wetlands. J. Hydrol. 296, 1-22.

Liaw, A., Wiener, M., 2002. Classification and regression by RandomForest. R News 2, 18-22.

Martin, V., Pfeiffer, D.U., Zhou, X., Xiao, X., Prosser, D.J., Guo, F., Gilbert, M., 2011. Spatial distribution and risk factors of highly pathogenic avian influenza (HPAI) H5N1 in China. PLoS Path. 7, e1001308.

Miller, O.D., Wilson, J.A., Ditchkoff, S.S., Lochmiller, R.L., 2000. Consumption of agricultural and natural foods by waterfowl migrating through central Oklahoma. Proc. Okla. Acad. Sci. 80, 25-31.

Moriguchi, S., Onuma, M., Goka, K., 2012. Potential risk map for avian influenza A virus invading Japan. Divers. Distrib.

Munster, V.J., Baas, C., Lexmond, P., Waldenstrom, J., Wallensten, A., Fransson, T., 2007. Spatial, temporal, and species variation in prevalence of influenza A viruses in wild migratory birds. PLoS Path. 3, 630-638.

Murphy, M.A., Evans, J.S., Storfer, A., 2010. Quantifying *Bufo boreas* connectivity in Yellowstone National Park with landscape genetics. Ecology 91, 252-261.

Nichols, J.D., Williams, B.K., 2006. Monitoring for conservation. Trends Ecol. Evol. 21, 668-673.

Nicodemus, K.K., 2011. Letter to the Editor: on the stability and ranking of predictors from random forest variable importance measures. Brief. Bioinform. 12, 369-373.

Northoff, E., 2008. Bird flu situation in Indonesia critical. FAO Newsroom http://www.fao.org/newsroom/en/news/2008/1000813/index.html.

Olson, D.M., Dinerstein, E., Wikramanaya, E.D., Burgess, N.D., Powell, G.V.N., Underwood, E.C., D'Amico, J.A., Itoua, I., Strand, H.E., Morrison, J.C., Loucks, C.J., Allnutt, T.F., Ricketts, T.H., Kura, Y., Lamoreux, J.F., Wettengel, W.W., Hedao, P., Kassem, K.R., 2001. Terrestrial Ecoregions of the World: A New Map of Life on Earth. Bioscience 51, 933-938.

Oyana, T.J., Dai, D., Scott, K.E., 2006. Spatiotemporal distributions of reported cases of the avian influenza H5N1 (bird flu) in southern China in early 2004. Avian Dis. 50, 508-515.

Parmley, E.J., Soos, C., Breault, A., Fortin, M., Jenkins, E., Kibenge, F., King, R., 2011. Detection of low pathogenic avian influenza viruses in wild ducks from Canada: comparison of two sampling methods. J. Wildl. Dis. 47, 466-470.

Parmley, J., Lair, S., Leighton, F.A., 2009. Canada's inter-agency wild bird influenza survey. Integr. Zool. 4, 409-417.

Pasick, J., Berhane, Y., Hisanaga, T., Kehler, H., Hooper-McGrevy, K., Handel, K., Neufeld, J., Argue, C., Leighton, F., 2010. Diagnostic test results and pathology associated with the 2007 Canadian H7N3 highly pathogenic avian influenza outbreak. Avian Dis. 54, 213-219.

Paul, M., Tavornpanich, S., Abrial, D., Gasqui, P., Charras-Garrido, M., Thanapongtharm, W., Xiao, X., Gilbert, M., Roger, F., Ducrot, C., 2010. Anthropogenic factors and the risk of highly pathogenic avian influenza H5N1: prospects from a spatial-based model. Vet. Res. 41, 28.

Pelzel, A., McCluskey, B., Scott, A., 2006. Review of the highly pathogenic avian influenza outbreak in Texas, 2004. J. Am. Vet. Med. Assoc. 228, 1869-1875.

Perkins, L., Swayne, D., 2001. Pathobiology of A/Chicken/Hong Kong/220/97 (H5N1) avian influenza virus in seven gallinaceous species. Vet. Pathol. 38, 149-164.

Pfeiffer, D.U., Minh, P.Q., Martin, V., Epprecht, M., Otte, M.J., 2007. An analysis of the spatial and temporal patterns of highly pathogenic avian influenza occurrence in Vietnam using national surveillance data. Vet. J. 174, 302-309.

Prasad, A.M., Iverson, L.R., Liaw, A., 2006. Newer classification and regression tree techniques: bagging and Random Forests for ecological prediction. Ecosystems 9, 181-199.

R Development Core Team, 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

Reemers, S.S., Leenen, D.v., Koerkamp, M.J.G., Haarlem, D.v., Haar, P.v.d., Eden, W.v., Vervelde, L., 2010. Early host responses to avian influenza A virus are prolonged and enhanced at transcriptional level depending on maturation of the immune system. Mol. Immunol. 47, 1675-1685.

Ruiz-Gazen, A., Villa, N., 2007. Storms prediction: logistic regression vs Random Forest for unbalanced data. Cas. Stud. Bus. Ind. Gov. Stat. 1, 91-101.

Sanderson, E.W., Jaiteh, M., Levy, M.A., Redford, K.H., Wannebo, A.V., Woolmer, G., 2002. The Human Footprint and the Last of the Wild. Bioscience 52, 891-904.

Sengupta, R., Rosenshein, L., Gilbert, M., Weiller, C., 2007. Ecoregional dominance in spatial distribution of avian influenza (H5N1) outbreaks. Emerg. Infect. Dis. 13, 1269-1271.

Seni, G., Elder, J., 2010. Ensemble methods in data mining: improving accuracy through combining predictions. Synth. Lect. Data Min. Knowl. Disc. 2, 1-126.

Senne, D., 2007. Avian Influenza in North and South America, 2002–2005. Avian Dis. 51, 167-173.

Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T., 2005. ROCR: visualizing classifier performance in R. Bioinf. Appl. Note 21, 3940-3941.

Spackman, E., Senne, D., Myers, T., Bulaga, L., Garber, L., Perdue, M., Lohman, K., Daum, L., Suarez, D., 2002. Development of a real-time reverse transcriptase PCR assay for type A influenza virus and the avian H5 and H7 hemagglutinin subtypes. J. Clin. Microbiol. 40, 3256-3260.

Stallknecht, D., Shane, S., 1988. Host range of avian influenza in free-living birds. Vet. Res. Commun. 12, 125-141.

Thomas, M.E., Bouma, A., Ekker, H.M., Fonken, A.J.M., Stegeman, J.A., Nielen, M., 2005. Risk factors for the introduction of high pathogenicity Avian Influenza virus into poultry farms during the epidemic in the Netherlands in 2003. Prev. Vet. Med. 69, 1-11.

Ward, M.P., Maftei, D., Apostu, C., Suru, A., 2008. Geostatistical visualisation and spatial statistics for evaluation of the dispersion of epidemic highly pathogenic avian influenza subtype H5N1. Vet. Res. 39, 22.

Williams, R.A.J., Fasina, F.O., Peterson, A.T., 2008. Predictable ecology and geography of avian influenza (H5N1) transmission in Nigeria and West Africa. Trans. R. Soc. Trop. Med. Hyg. 102, 471-479.

Zhou, K.H., O'Malley, J., Mauri, L., 2007. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation 115, 654-657.

CHAPTER 5

General Discussion

Avian influenza virus (AIV) is a disease that is nearly ubiquitous in wild populations of birds, particularly wild waterfowl and shorebirds. When it infects domestic poultry, chickens and turkeys in particular, the effects can be economically devastating. Furthermore, AIV has demonstrated the ability to infect a wide range of mammals including humans. Although human-to-human transmission is exceedingly rare, the mortality rate in humans from AIV contracted from birds exceeds 50%. Because AIV infects such a wide range of organisms, has the potential to cause pandemics in humans and other animals, and appears to be endemic to migratory waterfowl, it poses an intriguing and important challenge for data-mining and disease modeling. The highly pathogenic avian influenza (HPAI) strains of H5N1 are the focus of most predictive and risk modeling efforts; however, HPAI H5N1 strains in poultry make up but a small subset within the vast diversity of AIV. Wild birds are the reservoir of AIV and the origin of H5N1 and other HPAI outbreaks in domestic poultry. Thus, the primary contribution of this dissertation is that it addresses a gap in knowledge regarding the ecological niche of AIV in wild birds. The ecological niche described by these three projects presents a set of bioclimatic, geographic, and anthropogenic conditions that are correlated with AIV-positive cases in wild birds and differs from the niche of H5N1 HPAI. In addition, it outlines a methodology for studying environmental correlates of outbreak in wild birds. Future work will clarify the mechanistic role of the variables we found were correlated with AIV and improve the predictive accuracy of wild bird models.

Overview

In this thesis, I defined and modeled the ecological niche for the predicted occurrence of AIV in wild birds worldwide. Chapter 2 detailed the construction of the first Pacific Rim model of AIV in wild birds using a surveillance database from the Alaska Asia Avian Influenza Research (A3IR) project. This model was able to identify bioclimatic and anthropogenic characteristics of areas in which positive samples were collected. This information was then used to create a predictive map of AIV occurrence around the Pacific Rim. The most important predictor variables for this collection of data were low mean temperature in February (-19.5 °C), high temperature seasonality (a range of 142 °C), and a long distance from coast (800 km). These results are in contrast with previous AIV models, which found that high population density was an important predictor; however, the scope and the target of these models were very different from ours. Previous models restricted their sampling to warm, humid countries and were targeting only the HPAI H5N1 subtype in domestic poultry and/or humans. The current model made contributions to predictive modeling of AIV in the following ways: 1) at the time of its publication (2010), it was one of the largest models encompassing Mongolia, Japan, Russia and Alaska. A broad geographic scope is important in studying AIV because its host species can be highly mobile. 2) The target species of this model were wild birds. Not only is AIV enzootic to wild birds, but wild birds are the primary hosts that infect domestic poultry. Although wild birds have been identified as important factors in the transmission of AIV to domestic poultry [5, 9], few models have focused on wild birds and fewer still on strains and subtypes other than HPAI H5N1. 3) This model used

Random Forests, an ensemble data-mining algorithm that is well established in both business and science applications and, as this project demonstrated, can be applied to ecological data-mining questions as well.

Chapter 3 extended the findings of Chapter 2 by constructing an ecological niche model and predictive map of AIV in wild birds from a large, collaborative database of surveillance data by an international group of researchers. As of this writing, not only was it one of few niche models constructed for wild birds and low-pathogenicity avian influenza (LPAI), it was the first global-scale niche model for this system. This model described two niches for LPAI. The first was described as having low annual rainfall (<500 mm) and low temperatures (<10 °C in April and June, <-20 °C in November). The secondary niche had moderate annual rainfall (>1000 mm) and higher temperatures (> 28 °C). Temperature and humidity conditions favorable to AIV persistence in the environment were discussed as possible mechanisms underlying the characteristics of this global niche. This model contributed to the understanding of AIV by demonstrating how the usefulness of AIV surveillance data could be extended by using it for predictive modeling. In addition, this chapter focused on the dynamics of LPAI and underscored the importance of the largely neglected northern areas in the ecology of AIV.

Chapter 4 addressed highly imbalanced prevalence of AIV positivity in bird surveillance data. Imbalanced prevalence is a common condition in wildlife data where the condition researchers are looking for occurs at a low rate. Furthermore, prevalence is an important factor in model accuracy. Custom code was written to compare and evaluate the effectiveness of a balancing algorithm, a model selection algorithm, and a

combination of the two. Two databases from independent surveillance projects were also compared and contrasted in terms of the breadth of their surveillance efforts and its effects on model accuracy. Using a simple balancing algorithm and an under-sampling method, this chapter provided effective, real-world strategies for handling datasets with highly unbalanced prevalence to improve accuracy in predictive modeling. The highest-scoring model was used to construct a predictive map of the relative occurrence of LPAI around the Pacific Rim study area. Again, the greatest contiguous areas of high predicted occurrence were in northern areas, i.e. North and East Asia and Central Alaska. Unlike the previous models, areas of moderate risk were predicted in temperate and equatorial regions, some of which are consistent with HPAI H5N1 data. The niche described by the highest scoring model had needleleaf landcover, a mean temperature in April below freezing (-2.2 °C), a high temperature seasonality (143 °C), small distance from rivers, lakes, or wetlands (0 m), a low slope (0°), and a low human population density (0 persons/km^2).

Chapter 4 demonstrated that machine-learning and predictive modeling can be applied effectively to a system as complicated as AIV in wild birds. The algorithms examined in this chapter are applicable to many types of wildlife surveillance data when modeling almost any system with highly unbalanced prevalence. Furthermore, these models can be optimized, customized, implemented, and analyzed using R statistical programming language. R is a well-established language and free to use and we were able to build these models using publically accessible, online databases such as the

Influenza Research Database (IRD; used in Chapter 3) and the Canadian Cooperative Wildlife Health Centre influenza data (i.e. the CIWBI database in Chapter 4).
These initial studies into constructing ecological niche models by integrating machine-learning, predictive modeling, and GIS technologies can ultimately serve as an important tool by providing quantitative results on which to base decisions regarding the management of AIV in human and animal health. This chapter also discussed over-sampling of study sites in "traditional" surveillance, its detrimental effects on model accuracy, its diminishing returns in terms of usefulness and value over time, and suggests Adaptive Management principles to maximize limited resources and funding [11, 20].

The datasets used in this dissertation overlapped slightly because of the contribution of A3IR data. The 2005-2007 A3IR dataset shares 7% of its samples with the IRD dataset and 21% of its samples with the 2005-2007 A3IR dataset. The 2005-2010 10% of the 2005-2010 A3IR dataset comes from the 2005-2007 A3IR dataset and shares 6% of its samples with IRD. 2005-2010 A3IR contributes to 4% of the IRD dataset and 10% of this subset comes from 2005-2007 samples. The CIWBI dataset is independent from both the IRD and A3IR datasets. To reiterate discussion from Chapter 4, it is interesting to note that although A3IR and CIWBI have no samples in common, models constructed from these datasets performed relatively well against each other. I would predict that models constructed from CIWBI and UNIQUE would perform well against other datasets, such as the 2005-2007 A3IR and IRD. I would expect the IRD model to be a powerful predictor on other datasets due to the range of its predictor variables. Predictive modeling tends to rely on internal training and testing subsets, and while this

method is common and well-established, efforts should be made to apply models to outside datasets. Increased data-sharing would increase opportunities for this kind of evaluation.

The LPAI niche vs. the HPAI niche

All three ecological niche models used data from different sources: the original Alaska Asia A3IRB database, the NIH Influenza Research Database, the CIWBI live and dead bird surveillance database, and the updated A3IR database. The niche defined by all three models is consistent in that they identify cold regions. Two of the three models identified high temperature seasonality as an important predictor. The distance to the coastline or to water bodies was either not an important predictor or had a high partial dependence at a great distance from water, despite the predominance of waterfowl and shorebirds in these surveillance efforts, The maps of relative predicted occurrence all showed the largest contiguous patches of high scores in northern regions, especially in North and East Asia and Alaska.

Overall, these projects found that the niche of LPAI is different from that of HPAI H5N1. The risk factors for HPAI H5N1, as described by multiple models, include agricultural land use, especially aquaculture or rice cultivation; the presence of domestic poultry and waterfowl; high human impact (e.g. population density and associated anthropogenic characteristics); and topographical factors that encourage water pooling such as seasonal flooding or low slope angle. Nearly all the models covered small study areas in low latitude regions, which typically had strong anthropogenic predictors such as high human population density, agricultural land use, proximity to roads and cities, and

the presence of domestic poultry and ducks. A large part of the consistency between these models may be due to the fact that the models used a very limited number of predictors and based the selection of these predictors on those used by previous models. If no novel predictors are introduced, no novel results will be found. Data-mining has an advantage over some of these other machine-learning techniques because an algorithm like Random Forests can easily handle a large number of predictors and is relatively immune to noise or non-contributing predictors [2]. Many predictors can be conveniently tested thereby increasing the likelihood that novel correlates can be identified.

The major difference between our models and those for HPAI H5N1 is that northern regions are consistently identified as areas of high predicted occurrence. The importance of northern regions in the ecology of AIV may be based on two major factors: they possess environmental conditions that allow for viral persistence and they serve as summer breeding grounds for *Anatidae*, which make up the highest percentage of infected birds. Virus deposited one season could survive in the environment until the next season to infect immunologically-naïve, juvenile birds [26]. Furthermore, the high bird densities that occur during summer breeding enhance viral transmission through the fecal-oral route [32], especially by contaminated water [4, 14]. Studies have found that the highest rates of viral isolation in ducks come from juveniles during the summer breeding season [10, 12]. The higher prevalence of AIV in northern breeding grounds over southern wintering locations likely accounts for the north-south gradient in AIV prevalence [17]. It is important to note that the density plots of AIV-positivity on predictor variables lack strong correspondence between values where surveillance effort

was concentrated and where AIV-positive density was the highest. This suggests that locations of highest sampling effort do not necessarily yield a proportional amount of AIV-positive samples. Therefore, these predictive models are not simply species prediction maps for the bird hosts of AIV, i.e. the niche for AIV-positivity is distinct from collection effort.

While HPAI H5N1 is economically devastating and a concern for public health, I feel it is important to consider it in perspective of its place in the broad spectrum of all AIV, as well as in the context of its ecology and landscape. This subtype and strain does not arise spontaneously or exist in a vacuum; the dynamics of transmission are clearly linked to the wild bird LPAI source. The work described by this dissertation is an initial examination of the larger niche of AIV outbreak. Future research can clarify the mechanisms behind the correlated predictor variables and how they contribute to AIV-positivity. In turn, describing the mechanisms and identifying the drivers behind AIV-positivity will increase the accuracy of AIV outbreak prediction, which can then be applied to transmission to domestic poultry.

One aspect of the HPAI H5N1 models that would improve the understanding of the wild bird and domestic poultry interface is the clarification between two aspects of AIV outbreak: initial outbreak and transmission. After initial outbreak, the transmission between infected units (i.e. infected farms or villages) is clearly due to human-driven, contact between these units [21, 28, 31]. However, to study the role of wild birds, a distinction should be made between these transmission cases and the initial, emergent case or cases. While emergent cases can be identified through spatiotemporal analysis

[24], advanced phylogeographic techniques are powerful tools for the identification and isolation of the strain responsible for the emergent case. Furthermore, comparisons with other sequences could identify the species from which it originated, be it wild bird, swine, other mammal, or even human. In addition, these phylogeographic tools can provide a picture of the geographic spread of a strain, as well as the evolution and persistence of various lineages that emerge from the initial case [13]. A similar analysis in wild birds would be much more challenging due to difficulties in sampling from a single population when this population migrates thousands of miles on a seasonal basis. Theoretically, one could identify viral sequence from banded, re-captured individuals, but this would require a large number of re-captures and extensive data-sharing to accumulate sufficient data for analysis.

Technical aspects and software

This dissertation highlights the technological aspect of data-mining and predictive modeling. While Chapters 2 and 3 used licensed, commercial software for analysis, Chapter 4 compared this software with the results from free, open-access software. The licensed software (Salford Predictive Modeler) outperformed R in all cases, which brings into question the reliability of the widely used randomForest package. While randomForest is based on the original Fortran code published by Breiman and Cutler [2], Salford has the advantage of employing Dr. Adele Cutler (one of the original developers of the Random Forests algorithm) who has continued to develop Random Forests over the decades since the original code was published [6]. In addition, Salford Predictive

Modeler is the product of a software company and as such, we can assume it undergoes quality control and has resources for research and development.

The randomForest R package [15] was created shortly after Random Forests was published and has undergone limited evolution, although the author updates the package and is responsive to questions and suggestions (A. Liaw, pers. comm.). Salford states that Random Forests is not "black box" [23] and that its implementation is not opaque. The basis of this claim is that Predictive Modeler can produce a variety of analytical results. However, this does not address the fact that the specific algorithms being used to generate the trees are not readily accessible to the user. The finding that the Salford models have higher AUC values than their R counterparts and that they produce trees with a much smaller number of terminal nodes suggests fundamental differences between the two systems. However, the inner workings of the randomForest R package are opaque as well. Although the R code is accessible to the user, examination of this code reveals that it only serves as an interface that collects the user's data, packages it, and passes it to another language that performs the actual calculations. The randomForest is well established having been used extensively in published research, actively maintained, and updated over the past ten years. However, due to the open-access nature of R and R packages, this level of quality cannot be expected from all other packages. One major advantage of R is that it allows the user to write custom processes, employ new algorithms (such as with MIR and RSS), and automate testing. This customization is all but impossible with a software package such as Salford Predictive Modeler.

Future work

Now that it has been established that models can be built to predict AIV in wild birds on a large scale, future research can begin to address how LPAI is correlated with the predictor variables that describe its ecological niche. When the mechanics behind these relationships are understood, the driving factors behind AIV outbreak in wild birds will become better defined. This knowledge can be integrated with the extensive research that has already been done on HPAI H5N1. Understanding HPAI H5N1in the larger context of AIV will lead to better prediction and clearer identification of risk factors.

First, a clearer description of the LPAI niche could be produced with the addition of predictor variables. The models presented in this dissertation used similar sets of predictor variables due to the lack of high-resolution layers available on the scale required. I expect more predictor variables will become available in the future, further enhancing predictive modeling of AIV. Additional layers that would be particularly useful would include global agriculture, particularly cereal agriculture, which is an important source of forage for wild birds [16], and a layer indicating summer breeding grounds for waterfowl. Second, the surveillance data available for use in these models show spatiotemporal sampling bias. The current dogma states that waterfowl and shorebirds are the primary reservoir for AIV [1, 25], thus, surveillance efforts tend to be biased in favor of sampling these two orders. Sampling often occurs when waterfowl congregate in high densities in their summer breeding grounds and when the populations of AIV-susceptible juveniles [10, 12] are present; as a result, the data are temporally biased for the Northern Hemisphere summer and spatially biased toward arctic breeding

grounds. This bias may result in artificially high risk in northern regions. This is not to say all findings in this dissertation are spurious. It is clear that *Anatidae* play an important role in the ecology of AIV, whether or not they are the primary reservoir of AIV and the point of contact that results in HPAI outbreaks in domestic poultry. One interpretation could be that the strength of the signal in *Anatidae* is strong enough that northern regions remain important when all other species are treated as noise.

The Southern Hemisphere would be an interesting "testing ground" for the northern models. If these models were robust, they should predict important southern breeding grounds. One potential problem with this comparison is that there are no regions in the Southern Hemisphere comparable to the broad, boreal regions seen in Siberia and Alaska. In particular, when a density plot of latitude was made for the IRD data used in Chapter 3, the nature of the bifurcate niche became evident (Fig. 5.1). One AIV-positive peak appeared in high latitude regions (approximately 50°) as would be expected from sampling northern breeding grounds. A second strong peak appeared at approximately 10°. The tropical latitude of this second peak would correlate with the secondary niche described as being warm with low temperature seasonality. As this is the same range within which HPAI H5N1 models have been constructed, it may indicate an important tropical niche related to HPAI H5N1 transmission between wild birds and domestic poultry. Regardless, it would be valuable and interesting to model AIV with a greater contribution from Southern Hemisphere samples.

Finally, species bias toward ducks and shorebirds in sampling may diminish the role of passerines or small perching birds. It has been reported that more passerine

species are infected than waterfowl species [8]. While the prevalence of their AIV-positivity may be lower than that of waterfowl, passerines can be infected both naturally and in the laboratory, and shed live virus [19]. Breithaupt et al. (2010 [3]) found that experimentally infected passerines shed H5N1 primarily from the respiratory tract. Passerines are usually sampled from the cloaca or from their feces, which may account for the low levels of virus found in surveys. Although the small size of many passerines would make sampling a challenge, it would be useful to compare tracheal, cloacal, and fecal samples. If passerines shed virus primarily from the trachea, they still have the potential to transmit virus through shared drinking water [14, 27]. If indeed passerines are responsible for transmitting AIV to domestic poultry, then biosecurity issues become more complicated. While game birds and wild waterfowl are relatively large in body size, keeping small birds out of poultry housing could be challenging. One valuable project would be to model AIV risk in passerines and see if contact with wild waterfowl was an important factor. Understanding transmission between waterfowl and passerines would help to clarify their role as AIV host and possible source of infection of domestic poultry.

One way to alleviate some sampling bias would be through improved data sharing within the AIV research community. Bird capture, sample collection, and the associated laboratory testing are both time-consuming and labor intensive. If georeferenced surveillance data were readily accessible and made available through sites such as fludb.org (Influenza Research Database) it would clarify what species had been sampled, where, and when they had been sampled. This information would prevent redundant effort and provide data for predictive and spatiotemporal modeling. One important

contribution that models have for AIV study and surveillance is that they provide some level of quality control and error checking. As addressed in a previous section, models can highlight sampling bias, which in turn can direct future surveillance efforts. Even basic mapping can reveal errors in the data. All the databases used in this dissertation (with the exception of CIWBI) required extensive grooming and inspection of points with incorrect latitude/longitude coordinates. Without this error checking, some degree of noise would have been added to the models. Errors and extra work involved in data grooming could be avoided by simply looking at surveillance points on a map before submitting them to a database.

The lack of some central repository for surveillance data is not unique to LPAI. Interestingly, it is surprisingly difficult to find a comprehensive map or list of countries affected by HPAI H5N1. The World Health Organization maintains the information on human cases [33]. The United States Department of Agriculture Animal and Plant Health Inspection Service has a webpage listing countries and regions affected by HPAI [29] (it is not specified whether these outbreaks were in domestic or wild animals); however, Europe, which has had outbreaks [1, 22] is missing from this list. The United States Geologic Survey National Wildlife Health Center (USGS/NWHC) website has an interactive map [30] that has filtering functions for HPAI, LPAI, non-H5N1 in humans, domestic animals, or wildlife. A search of all HPAI H5N1 is notably missing cases in China, Nigeria, and Europe. The Food and Agriculture Organization of the United Nations Global Animal Disease Information System (FAO/EMPRES) [7] maps recent outbreaks reported to the FAO, but only for the past six months. Despite the potential

economic losses to the poultry industry and impact on human health, it is surprising that no agency appears to present a comprehensive overview of the disease.

Surveillance and Adaptive Management principles

We hope that AIV and wild bird surveillance philosophy will shift in favor of collaborative efforts that provide useful data for predictive modeling, rather than focusing all collection activities on the same locations year after year. Surveillance guided by hypothesis testing requires innovative thinking and results in better data for better management decisions. Using such an Adaptive Management feedback loop ensures that surveillance and research methods are focused on answering specific research questions, such that the component modeling can serve as quality control by identifying areas or ranges of variables that require additional sampling. The construction of models and maps of predicted relative occurrence are just one step in an iterative cycle of hypothesis development, testing, re-development of hypotheses, and re-testing. Ideally, collection in the field and laboratory testing would occur in cooperation with modeling, machine-learning, and data-mining projects. This arrangement would be mutually beneficial to computational biology, wildlife biology, and laboratory science. Based on existing surveillance data, researchers would construct predictive or spatiotemporal models. These models and machine-learning results could guide future field and laboratory studies to fill apparent gaps in available data and suggest new hypotheses for collection and testing. These collected data would in turn strengthen existing models and provide fresh data to guide new modeling and machine-learning projects, which in turn would produce results

for future field and laboratory studies. With each cycle of this collaborative system, the data produced becomes more accurate and more useful.

Due to the global scale of the disease, AIV research will benefit from open access data sharing policies such as those put forth by NIH and NSF. As stated on the NIH Data Sharing Policy webpage "Data sharing is essential for expedited translation of research results into knowledge, products and procedures to improve human health." [18]. In collecting data from migratory waterfowl that travel long distances yearly, data sharing is crucial. It will maximize surveillance effort and resources by preventing redundant sampling. Targeted surveillance, collaborative research, and open access data would benefit the researchers as well as those who would use these data for disease management, policy development, and decision making for the benefit of public health and health policy for humans and animals.

FIGURES

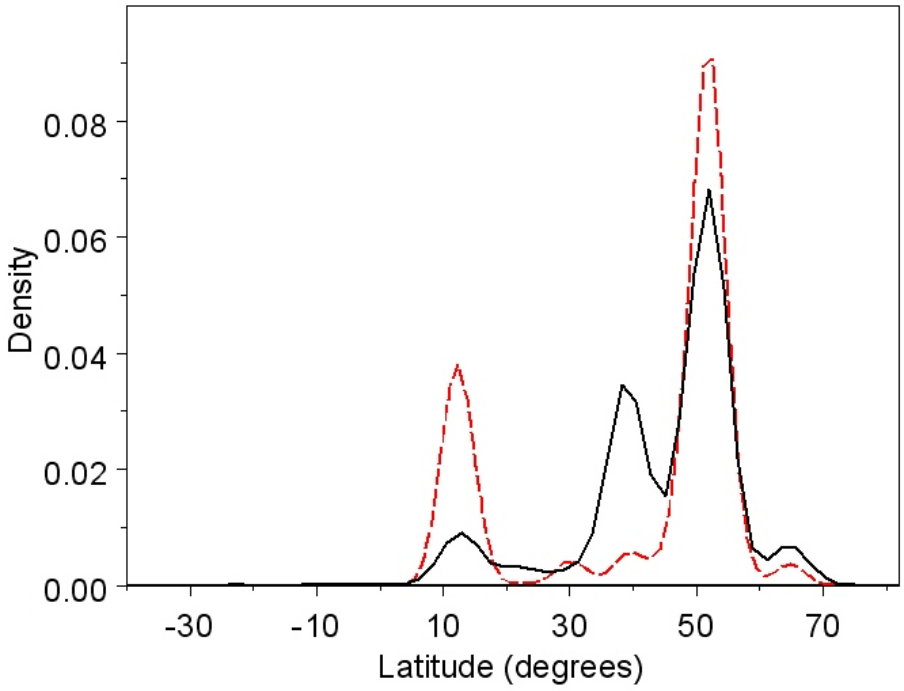

Figure 5.1. Density plot of latitude. This figure shows the range of values over which sampling occurred in the subset from the Influenza Research Database that was used to construct a global model of avian influenza in Chapter 3. The black line represents where AIV-negative samples were collected and the dotted red line represents where AIV-positive samples were found. The AIV-positive samples display a bifurcate niche at temperate and tropical latitudes.

LITERATURE CITED

[1] Alexander DJ: **A review of avian influenza in different bird species** *Vet Microbiol* 2000, **74**:3-13.

[2] Breiman L: **Random Forests** *Mach Learn* 2001, **45**:5-32.

[3] Breithaupt A, Kalthoff D, Dale J, Bairlein F, Beer M, Teifke J: **Neurotropism in Blackcaps (*Sylvia atricapilla*) and red-billed queleas (*Quelea quelea*) after highly pathogenic avian influenza virus H5N1 infection** *Vet Pathol* 2010, **48**:924-932.

[4] Brown JD, Goekjian G, Poulson R, Valeika S, Stallknecht DE: **Avian influenza virus in water: infectivity is dependent on pH, salinity and temperature** *Vet Microbiol* 2007, **136**:20-26.

[5] Cecchi G, Ilemobade A, Brun YL, Hogerwerf L, Slingenbergh J: **Agro-ecological features of the introduction and spread of the highly pathogenic avian influenza (HPAI) H5N1 in northern Nigeria** *Geospat Health* 2008, **3**:7-16.

[6] Cutler DR, Edwards TC, Jr., Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ: **Random Forests for classification in ecology** *Ecology* 2007, **88**:2783-2792.

[7] Food and Agriculture Organization of the United Nations Global Animal Disease Information System (FAO/EMPRES): Disease Events Map - H5N1 HPAI [http://empres-i.fao.org/eipws3g/#h=1].

[8] Fuller TL, Saatchi SS, Curd EE, Toffelmeier E, Thomassen HA, Buermann W, Smith TB: **Mapping the risk of avian influenza in wild birds in the U.S.** *BMC Infect Dis* 2010, **10**:187.

[9] Gilbert M, Xiao X, Domenech J, Lubroth J, Martin V, Slingenbergh J: **Anatidae migration in the western palearctic and spread of highly pathogenic avian influenza H5N1 virus** *Emerg Infect Dis* 2006, **12**:1650-1656.

[10] Hanson BA, Stallknecht DE, Swayne DE, Lewis LA, Senne DA: **Avian influenza viruses in Minnesota ducks during 1998-2000** *Avian Dis* 2003, **47**:867-871.

[11] Huettmann F: **Modern adaptive management: adding digital opportunities towards a sustainable world with new values** *Forum on Public Policy: Climate Change and Sustainable Development* 2007, **3**:337-342.

[12] Krauss S, Walker D, Pryor P, Niles L, Chenghong L, Hinshaw VS, Webster RG: **Influenza A viruses of migrating wild aquatic birds in North America** *Vector Borne Zoonotic Dis* 2004, **4**:177-189.

[13] Lam TT-Y, Hon C-C, Lemey P, Pybus OG, Shi M, Tun HM, Li J, Jiang J, Holmes EC, Leung FC-C: **Phylodynamics of H5N1 avian influenza virus in Indonesia** *Mol Ecol* 2012, **21**:3062-3077.

[14] Leung YHC, Zhang L-J, Chow C-K, Tsang C-L, Ng C-F, Wong C-K, Guan Y, Peiris JSM: **Poultry drinking water used for avian influenza surveillance** *Emerging Infectious Diseases [serial on the internet]* 2007:Available from http://www.cdc.gov/EID/content/13/19/1380.htm.

[15] Liaw A, Wiener M: **Classification and regression by RandomForest** *R News* 2002, **2**:18-22.

[16] Miller OD, Wilson JA, Ditchkoff SS, Lochmiller RL: **Consumption of agricultural and natural foods by waterfowl migrating through central Oklahoma** *Proc Okla Acad Sci* 2000, **80**:25-31.

[17] Munster VJ, Baas C, Lexmond P, Waldenstrom J, Wallensten A, Fransson T: **Spatial, temporal, and species variation in prevalence of influenza A viruses in wild migratory birds** *PLoS Path* 2007, **3**:630-638.

[18] National Insitute of Health: NIH Data Sharing Policy [http://grants.nih.gov/grants/policy/data_sharing/].

[19] Nestorowicz A, Kawaoka Y, Bean W, Webster R: **Molecular analysis of the hemagglutinin genes of Australian H7N7 influenza viruses: role of passerine birds in maintenance or transmission?** *Virology* 1987, **160**:411-418.

[20] Nichols JD, Williams BK: **Monitoring for conservation** *Trends Ecol Evol* 2006, **21**:668-673.

[21] Paul M, Tavornpanich S, Abrial D, Gasqui P, Charras-Garrido M, Thanapongtharm W, Xiao X, Gilbert M, Roger F, Ducrot C: **Anthropogenic factors and the risk of highly pathogenic avian influenza H5N1: prospects from a spatial-based model** *Vet Res* 2010, **41**:28.

[22] Reperant LA, Fučkar NS, Osterhaus ADME, Dobson AP, Kuiken T: **Spatial and Temporal Association of Outbreaks of H5N1 Influenza Virus Infection in Wild Birds with the 0C Isotherm** *PLoS Path* 2010, **6**:e1000854.

[23] Salford Systems: Is RandomForests a black box? [https://www.salford-systems.com/en/products/randomforests/faqs/item/136-is-randomforests-a-black-box?].

[24] Souris M, Gonzalez J-P, Shanmugasundaram J, Corvest V, Kittayapong P: **Retrospective space-time analysis of H5N1 Avian Influenza emergence in Thailand** *Int J Health Geogr* 2010, **9**.

[25] Stallknecht D, Shane S: **Host range of avian influenza in free-living birds** *Vet Res Commun* 1988, **12**:125-141.

[26] Stallknecht DE, Goekjian VH, Wilcox BR, Poulson RL, Brown JD: **Avian influenza virus in aquatic habitats: what do we need to learn** *Avian Dis* 2010, **54**:461-465.

[27] Sturm-Ramirez KM, Hulse-Post DJ, Govorkova EA, Humberd J, Seiler P, Puthavathana P, Buranathai C, Webster RG: **Are ducks contributing to the endemicity of highly pathogenic H5N1 influenza virus in Asia?** *J Virol* 2005, **79**:11269-11279.

[28] Thomas ME, Bouma A, Ekker HM, Fonken AJM, Stegeman JA, Nielen M: **Risk factors for the introduction of high pathogenicity Avian Influenza virus into poultry farms during the epidemic in the Netherlands in 2003** *Prev Vet Med* 2005, **69**:1-11.

[29] United States Department of Agriculture Animal and Plant Health Inspection Service (USDA/APHIS): Countries/regions affected by highly pathogenic avian influenza [http://www.aphis.usda.gov/import_export/animals/animal_import/animal_imports_hpai.shtml].

[30] United States Geologic Survey National Wildlife Health Center (USGS/NWHC): Avian Influenza Map [http://www.nwhc.usgs.gov/map/].

[31] Ward MP, Maftei D, Apostu C, Suru A: **Geostatistical visualisation and spatial statistics for evaluation of the dispersion of epidemic highly pathogenic avian influenza subtype H5N1** *Vet Res* 2008, **39**:22.

[32] Webster RG, Yakhno M, Hinshaw VS, Bean WJ, Murti KG: **Intestinal influenza: replication and characterization of influenza viruses in ducks** *Virology* 1978, **84**:268-278.

[33] World Health Organization: Cumulative number of confirmed human cases for avian influenza A (H5N1) reported to WHO, 2003-2012 [http://www.who.int/influenza/human_animal_interface/EN_GIP_20120810CumulativeNumberH5N1cases.pdf].

APPENDICES

Appendix A. List of bird species in the Alaska Asia Avian Influenza Research 2005-2007 database. The database was used to construct the Pacific Rim model (Chapter 2) and this list includes order, scientific name, total number of samples collected for each species (# Samples) , and number of samples testing positive for avian influenza virus (# AIV Pos) for each species. This dataset contain 269 species of bird.

Order	Genus	Species	# Samples	# AIV Pos
Accipitriformes	*Accipiter*	*spp.*	4	0
	Accipiter	*spp.*	145	0
	Accipiter	*striatus*	5	0
	Aquila	*chrysaetos*	6	0
	Aquila	*clanga*	1	0
	Aquila	*nipalensis*	2	0
	Aquila	*pennata*	1	0
Anseriformes	*Anas*	*acuta*	1537	258
	Anas	*americana*	428	9
	Anas	*clypeata*	170	36
	Anas	*crecca*	152	30
	Anas	*platyrhynchos*	917	228
	Anser	*albifrons*	1663	3
	Anser	*indicus*	83	0
	Aythya	*affinis*	446	10
	Aythya	*affinis*	446	10
	Aythya	*collaris*	21	1
	Aythya	*marila*	9	1
	Aythya	*spp.*	4	0
	Aythya	*valisineria*	33	0
	Branta	*bernicla nigricans*	2391	14
	Branta	*canadiensis occidentalis*	197	14
	Branta	*canadensis*	5	0
	Bucephala	*albeola*	34	0
	Bucephala	*clangula*	5	0
	Bucephala	*islandica*	45	1
	Cygnus	*buccinator*	2	0
	Cygnus	*cygnus*	1	0
	Histrionicus	*histrionicus*	173	4
	Melanitta	*deglandi*	1	0
	Mergus	*merganser*	69	0
	Mergus	*serrator*	1	0
	Somateria	*spectabilis*	51	0
Caprimulgiformes	*Caprimulgus*	*europaeus*	23	0
Charadriiformes	*Actitis*	*hypoleucos*	8	0
	Actitis	*macularius*	1	0
	Aethia	*cristatella*	8	0

Appendix A continued

Order	Genus	Species	# Samples	# AIV Pos
Charadriiformes	*Aethia*	*pusilla*	8	0
	Arenaria	*interpres*	1	0
	Arenaria	*spp.*	2	0
	Calidris	*alpina*	11	0
	Calidris	*canutus*	1	0
	Calidris	*mauri*	54	1
	Calidris	*melanotos*	2	0
	Calidris	*pusilla*	29	0
	Calidris	*ruficollis*	111	0
	Calidris	*subminuta*	7	0
	Charadrius	*dubius*	1	0
	Charadrius	*mongolus*	12	0
	Chroicocephalus	*philadelphia*	1	0
	Gallinago	*gallinago*	1	0
	Gallinago	*hardwickii*	3	0
	Gallinago	*hardwickii*	75	0
	Gallinago	*spp.*	5	0
	Larus	*canus*	2	0
	Larus	*canus*	34	1
	Larus	*crassirostris*	296	0
	Larus	*ridibundus*	2	0
	Larus	*schistisagus*	906	0
	Larus	*spp.*	68	0
	Limicola	*falcinellus*	1	0
	Limosa	*lapponica*	1	0
	Limosa	*limosa*	1	0
	Onychoprion	*aleuticus*	2	0
	Phalaropus	*lobatus*	50	0
	Philomachus	*pugnax*	6	0
	Rissa	*tridactyla*	21	0
	Tringa	*brevipes*	100	0
	Tringa	*flavipes*	3	0
	Tringa	*incana*	8	0
	Tringa	*nebularia*	16	0
	Tringa	*solitaria*	8	0
	Uria	*aalge*	10	0
	Uria	*lomvia*	8	0
	Xenus	*cinereus*	2	0
Columbiformes	*Columba*	*eversmanni*	1	0
	Columba	*oenas*	83	0
	Columba	*palumbus*	5	0
	Columbiforme	*spp.*	24	0
	Treron	*sieboldii*	67	0

Appendix A continued

Order	Genus	Species	# Samples	# AIV Pos
Coraciiformes	*Alcedo*	*atthis*	1	0
	Alcedo	*atthis*	1	0
	Merops	*apiaster*	12	0
Cuculiformes	*Cuculidae*	*spp.*	2	0
Falconiformes	*Buteo*	*buteo*	7	0
	Buteo	*buteo*	2	0
	Buteo	*hemilasius*	3	0
	Buteo	*rufinus*	6	0
	Circus	*cyaneus*	1	0
	Circus	*cyaneus*	2	0
	Circus	*macrourus*	6	0
	Circus	*pygargus*	1	0
	Circus	*spp.*	2	0
	Falco	*cherrug*	30	0
	Falco	*columbarius*	1	0
	Falco	*naumanni*	21	0
	Falco	*rusticolus*	69	0
	Falco	*sparverius*	23	2
	Falco	*spp.*	30	0
	Falco	*spp.*	2	0
	Falco	*tinnunculus*	6	0
	Haliaeetus	*albicilla*	1	0
	Haliaeetus	*leucocephalus*	1	0
	Haliaeetus	*pelagicus*	1	0
	Milvus	*migrans*	15	0
	Pernis	*apivorus*	1	0
Gruiformes	*Anthropoides*	*virgo*	1	0
	Otis	*tarda*	1	0
Passeriformes	*Acrocephalus*	*bistrigiceps*	1	0
	Acrocephalus	*dumetorum*	4	0
	Acrocephalus	*spp.*	1	0
	Alauda	*gulgula*	1	0
	Anthus	*gustavi*	4	0
	Anthus	*hodgsoni*	10	0
	Anthus	*trivialis*	1	0
	Anthus	*trivialis*	7	0
	Bombycilla	*garrulus*	4	0
	Bradypterus	*tacsanowskius*	1	0
	Calandrella	*spp.*	1	0
	Calcarius	*lapponicus*	6	0
	Carduelis	*cannabina*	1	0
	Carduelis	*flammea*	429	1
	Carduelis	*hornemanni*	65	0

Appendix A continued

Order	Genus	Species	# Samples	# AIV Pos
Passeriformes	*Carduelis*	*pinus*	1	0
	Carduelis	*sinica*	3	0
	Carduelis	*sinica*	2	0
	Carduelis	*spinus*	12	0
	Carpodacus	*erythrinus*	2	0
	Carpodacus	*erythrinus*	5	0
	Catharus	*guttatus*	70	2
	Catharus	*minimus*	176	4
	Catharus	*ustulatus*	266	10
	Cecropis	*daurica*	1	0
	Certhia	*americana*	1	0
	Cettia	*spp.*	4	0
	Colluricincla	*harmonica*	7	0
	Corvus	*corone*	2	0
	Corvus	*corone cornix*	1	0
	Corvus	*frugilegus*	161	0
	Corvus	*macrorhynchos*	5	0
	Corvus	*monedula*	49	0
	Corvus	*ruficollis*	1	0
	Corvus	*spp.*	42	0
	Cyanistes	*flavipectus*	2	0
	Cyanopica	*cyanus*	7	0
	Dendroica	*petechia*	264	0
	Dendroica	*striata*	45	1
	Dendroica	*townsendi*	10	0
	Emberiza	*aureola*	12	0
	Emberiza	*buchanani*	1	0
	Emberiza	*hortulana*	3	0
	Emberiza	*leucocephalos*	3	0
	Emberiza	*pusilla*	10	0
	Emberiza	*rustica*	15	0
	Emberiza	*schoeniclus*	36	0
	Emberiza	*spodocephala*	234	0
	Empidonax	*alnorum*	28	0
	Empidonax	*hammondii*	121	1
	Erithacus	*akahige*	1	0
	Euphagus	*carolinus*	24	0
	Ficedula	*parva*	23	0
	Fringilla	*coelebs*	12	0
	Fringilla	*montifringilla*	6	0
	Garrulus	*glandarius*	7	0
	Hirundo	*rustica*	195	0
	Ixoreus	*naevius*	7	0

Appendix A continued

Order	Genus	Species	# Samples	# AIV Pos
Passeriformes	*Junco*	*hyemalis*	2073	105
	Junco	*hyemalis*	379	0
	Lanius	*bucephalus*	1	0
	Lanius	*excubitor*	4	0
	Locustella	*certhiola*	1	0
	Locustella	*lanceolata*	6	0
	Locustella	*ochotensis*	86	0
	Locustella	*seebohmi*	1	0
	Loxia	*curvirostra*	2	0
	Luscinia	*calliope*	143	0
	Luscinia	*cyane*	3	0
	Luscinia	*luscinia*	1	0
	Luscinia	*svecica*	5	0
	Melospiza	*lincolnii*	516	6
	Microscelis	*amaurotis*	2	0
	Miliaria	*calandra*	1	0
	Motacilla	*alba*	20	0
	Motacilla	*alba ocularis*	20	0
	Motacilla	*lugens*	1	0
	Motacilla	*personata*	2	0
	Motacilla	*spp.*	11	0
	Motacilla	*tschutschensis*	229	0
	Muscicapa	*dauurica*	14	0
	Oenanthe	*pleschanka*	1	0
	Oreothlypis	*celata*	928	14
	Oreothlypis	*peregrina*	1	0
	Parkesia	*noveboracensis*	204	1
	Parus	*bokharensis*	8	0
	Parus	*major*	31	0
	Passer	*domesticus*	16	0
	Passer	*domesticus*	2	0
	Passer	*hispaniolensis*	14	0
	Passer	*montanus*	24	0
	Passer	*rutilans*	8	0
	Passerculus	*sandwichensis*	128	0
	Passerella	*iliaca*	126	2
	Periparus	*ater*	2	0
	Perisoreus	*canadensis*	5	0
	Phoenicurus	*auroreus*	166	0
	Phoenicurus	*ochruros*	1	0
	Phylloscopus	*borealis*	349	0
	Phylloscopus	*borealoides*	16	0
	Phylloscopus	*coronatus*	15	0

Appendix A continued

Order	Genus	Species	# Samples	# AIV Pos
Passeriformes	*Phylloscopus*	*humei*	1	0
	Phylloscopus	*inornatus*	14	0
	Phylloscopus	*proregulus*	44	0
	Phylloscopus	*schwarzi*	3	0
	Phylloscopus	*spp.*	158	0
	Pica	*pica*	1	0
	Poecile	*atricapillus*	228	2
	Poecile	*hudsonica*	14	0
	Poecile	*montanus*	11	0
	Poecile	*palustris*	3	0
	Pyrrhula	*murina*	2	0
	Regulus	*calendula*	87	7
	Riparia	*riparia*	18	0
	Riparia	*riparia diluta*	25	0
	Saxicola	*rubicola*	3	0
	Setophaga	*coronata coronata*	1007	11
	Sitta	*europaea*	7	0
	Sitta	*spp.*	1	0
	Sitta	*spp.*	13	0
	Spizella	*arborea*	801	7
	Spizella	*arborea*	33	0
	Spizella	*passerina*	1	0
	Sturnus	*cineraceus*	28	0
	Sylvia	*communis*	3	0
	Sylvia	*curruca*	12	0
	Tachycineta	*bicolor*	1	0
	Tarsiger	*cyanurus*	16	0
	Turdus	*atrogularis*	3	0
	Turdus	*cardis*	5	0
	Turdus	*chrysolaus*	2	0
	Turdus	*eunomus*	5	0
	Turdus	*merula*	2	0
	Turdus	*migratorius*	171	1
	Turdus	*naumanni*	5	0
	Turdus	*obscurus*	4	0
	Turdus	*spp.*	26	0
	Uragus	*sibiricus*	5	0
	Vermivora	*celata*	464	7
	Wilsonia	*pusilla*	128	0
	Zonotrichia	*atricapilla*	17	0
	Zonotrichia	*leucophry*	204	0
	Zoothera	*sibirica*	1	0
	Zosterops	*japonicus*	3	0

Appendix A continued

Order	Genus	Species	# Samples	# AIV Pos
Piciformes	*Colaptes*	*auratus*	1	0
	Colaptes	*auratus*	1	0
	Dendrocopos	*leucotos*	2	0
	Dendrocopos	*minor*	3	0
	Dryocopus	*martius*	1	0
	Jynx	*spp.*	1	0
	Picoides	*arcticus*	1	0
	Picoides	*dorsalis*	2	0
	Picoides	*pubescens*	8	0
	Picoides	*villosus*	1	0
	Picus	*canus*	1	0
Procellariiformes	*Oceanodroma*	*leucorhoa*	4	0
Strigiformes	*Aegolius*	*funereus*	109	0
	Asio	*otus*	1	0
	Bubo	*blakistoni*	2	0
	Otus	*scops*	8	0
Suliformes	*Phalacrocorax*	*auritus*	132	1

Appendix B. List of bird species from the NIH Influenza Research Database (IRD). This subset of the IRD database was used to construct the Global Model (Chapter 3) and this list includes order, scientific name, total number of samples collected for each species (# Samples), and number of samples testing positive for avian influenza virus (# AIV Pos) for each species. This dataset contains 396 species of bird.

Order	Genus	Species	# Samples	# AIV Pos
Accipitriformes	*Accipiter*	*badius*	1	0
	Accipiter	*cooperii*	62	0
	Accipiter	*gularis*	1	0
	Aquila	*heliaca*	1	0
	Cathartes	*aura*	28	0
	Spizaetus	*cirrhatus*	1	0
Anseriformes	*Aix*	*galericulata*	25	0
	Aix	*sponsa*	639	0
	Alopochen	*aegyptiacus*	135	2
	Amazonetta	*brasiliensis*	4	0
	Anas	*acuta*	2246	28
	Anas	*americana*	1622	14
	Anas	*bahamensis*	1	0
	Anas	*bahamensis*	3	0
	Anas	*carolinensis*	173	12
	Anas	*clypeata*	1859	58
	Anas	*crecca*	2763	18
	Anas	*crecca carolinensis*	146	0
	Anas	*cyanoptera*	249	3
	Anas	*discors*	1192	86
	Anas	*falcata*	8	0
	Anas	*formosa*	2	0
	Anas	*penelope*	1837	50
	Anas	*platyrhynchos*	18478	736
	Anas	*platyrhynchos/rubripes*	3	0
	Anas	*platyrhynchos/strepera*	1	0
	Anas	*poecilorhyncha*	4	0
	Anas	*querquedula*	5	0
	Anas	*rubripes*	6	1
	Anas	*strepera*	1649	22
	Anser	*albifrons*	5471	145
	Anser	*anser*	509	2
	Anser	*brachyrhynchus*	3	0
	Anser	*caerulescens*	162	0
	Anser	*cygnoides*	168	29
	Anser	*fabalis*	912	25
	Anser	*indicus*	442	29
	Aythya	*affinis*	487	0
	Aythya	*americana*	24	5

Appendix B continued

Order	Genus	Species	# Samples	# AIV Pos
Anseriformes	*Aythya*	*collaris*	399	8
	Aythya	*ferina*	3	0
	Aythya	*fuligula*	10	0
	Aythya	*marila*	12	0
	Aythya	*nyroca*	1	0
	Aythya	*valisineria*	134	0
	Branta	*bernicla*	20	0
	Branta	*canadensis*	101	0
	Branta	*hutchinsii*	28	0
	Branta	*leucopsis*	514	8
	Bucephala	*albeola*	152	3
	Bucephala	*clangula*	47	0
	Bucephala	*islandica*	2	0
	Cairina	*moschata*	227	0
	Chen	*caerulescens*	120	0
	Chen	*rossii*	36	2
	Cygnus	*buccinator*	19	0
	Cygnus	*columbianus*	11	0
	Cygnus	*cygnus*	226	23
	Cygnus	*olor*	893	0
	Dendrocygna	*autumnalis*	44	0
	Dendrocygna	*javanica*	416	43
	Dendrocygna	*viduata*	19	0
	Fuligula	*affinis*	23	0
	Lophodytes	*cucullatus*	13	0
	Melanitta	*fusca*	4	0
	Melanitta	*perspicillata*	12	0
	Mergellus	*albellus*	5	0
	Mergus	*merganser*	22	0
	Mergus	*serrator*	1	0
	Netta	*erythrophthalma*	1	0
	Netta	*rufina*	3	0
	Nettapus	*coromandelianus*	5	1
	Oxyura	*jamaicensis*	73	0
	Oxyura	*leucocephala*	4	0
	Tadorna	*ferruginea*	187	16
	Tadorna	*tadorna*	15	2
Apodiformes	*Cypsiurus*	*balasiensis*	3	1
Caprimulgiformes	*Caprimulgus*	*macrurus*	2	0
Charadriiformes	*Actitis*	*hypoleucos*	97	1
	Arenaria	*interpres*	286	7
	Calidris	*acuminata*	3	1
	Calidris	*alba*	53	0

Appendix B continued

Order	Genus	Species	# Samples	# AIV Pos
Charadriiformes	*Calidris*	*alpina*	428	4
	Calidris	*canutus*	16	0
	Calidris	*ferruginea*	15	0
	Calidris	*mauri*	169	0
	Calidris	*minuta*	14	1
	Calidris	*minutilla*	45	0
	Calidris	*pusilla*	83	0
	Calidris	*ruficollis*	16	0
	Calidris	*subminuta*	58	7
	Calidris	*temminckii*	14	2
	Cerorhinca	*monocerata*	2	0
	Charadrius	*alexandrinus*	12	4
	Charadrius	*dubius*	159	4
	Charadrius	*hiaticula*	19	0
	Charadrius	*semipalmatus*	2	0
	Charadrius	*vociferus*	1	0
	Chlidonias	*leucopterus*	4	1
	Gallinago	*gallinago*	147	2
	Gallinago	*media*	1	0
	Gallinago	*spp.*	12	1
	Gallinago	*stenura*	7	0
	Glareola	*lactea*	2	0
	Haematopus	*ostralegus*	400	0
	Himantopus	*himantopus*	1	0
	Hydrophasianus	*chirurgus*	19	2
	Larus	*argentatus*	307	14
	Larus	*argentatus cachinnans*	134	0
	Larus	*armenicus*	64	0
	Larus	*atricilla*	23	2
	Larus	*cachinnans*	4	0
	Larus	*californicus*	186	0
	Larus	*canus*	1256	0
	Larus	*delawarensis*	5	0
	Larus	*fuscus*	14	0
	Larus	*marinus*	2	1
	Larus	*melanocephalus*	8	0
	Larus	*minutus*	2	0
	Larus	*occidentalis*	10	0
	Larus	*relictus*	2	0
	Larus	*ridibundus*	2824	84
	Limicola	*falcinellus*	9	0
	Limnodromus	*griseus*	29	0
	Limnodromus	*scolopaceus*	49	0

Appendix B continued

Order	Genus	Species	# Samples	# AIV Pos
Charadriiformes	*Limnodromus*	*semipalmatus*	1	0
	Limosa	*lapponica*	2	0
	Metopidius	*indicus*	19	0
	Numenius	*arquata*	6	0
	Phalaropus	*lobatus*	4	0
	Philomachus	*pugnax*	30	0
	Pluvialis	*apricaria*	4	0
	Pluvialis	*fulva*	39	2
	Pluvialis	*squatarola*	20	0
	Rostratula	*benghalensis*	61	14
	Rynchops	*niger*	3	0
	Scolopax	*rusticola*	7	0
	Stercorarius	*longicaudus*	1	0
	Sterna	*forsteri*	260	0
	Sterna	*hirundo*	13	0
	Tringa	*erythropus*	7	0
	Tringa	*glareola*	359	4
	Tringa	*melanoleuca*	1	0
	Tringa	*nebularia*	7	0
	Tringa	*ochropus*	6	1
	Tringa	*solitaria*	1	0
	Tringa	*stagnatilis*	3	0
	Tringa	*totanus*	69	3
	Turnix	*suscitator*	12	1
	Turnix	*sylvatica*	19	10
	Uria	*aalge*	68	0
	Vanellus	*chilensis*	1	0
	Vanellus	*vanellus*	5	0
	Xenus	*cinereus*	18	0
Ciconiiformes	*Anastomus*	*oscitans*	26	0
	Bubulcus	*ibis*	64	3
	Ciconia	*ciconia*	110	0
Columbiformes	*Columba*	*livia*	761	3
	Columba	*palumbus*	1	0
	Columbia	*livia*	1	0
	Columbina	*minuta*	5	0
	Columbina	*passerina*	1	0
	Columbina	*talpacoti*	33	2
	Geopelia	*cuneata*	2	0
	Geopelia	*striata*	37	4
	Leptotila	*plumbeiceps*	4	0
	Streptopelia	*chinensis*	56	3
	Streptopelia	*orientalis*	2	0

Appendix B continued

Order	Genus	Species	# Samples	# AIV Pos
Columbiformes	*Streptopelia*	*risoria*	1	0
	Streptopelia	*tranquebarica*	80	12
	Zenaida	*asiatica mearnsi*	1	0
Coraciiformes	*Alcedo*	*atthis*	37	3
	Alcedo	*meninting*	1	0
	Halcyon	*pileata*	1	0
	Halcyon	*smyrnensis*	1	0
	Megaceryle	*alcyon*	2	1
	Merops	*orientalis*	8	0
	Merops	*philippinus*	8	0
	Upupa	*epops*	1	0
Cuculiformes	*Cacomantis*	*merulinus*	4	0
	Centropus	*sinensis*	6	1
	Crotophaga	*sulcirostris*	12	0
	Phaenicophaeus	*tristis*	1	0
	Piaya	*cayana*	2	0
Falconiformes	*Buteo*	*buteo*	3	0
	Coragyps	*atratus*	6	0
	Falco	*peregrinus*	8	0
	Gyps	*bengalensis*	2	0
	Haliaeetus	*leucocephalus*	81	1
	Haliastur	*indus*	1	0
Galliformes	*Arborophila*	*chloropus*	1	0
	Coturnix	*chinensis*	12	2
	Coturnix	*coturnix*	7	0
	Coturnix	*japonica*	1	0
	Francolinus	*pintadeanu*	1	0
	Meleagris	*gallopavo*	7	0
	Numida	*meleagris*	72	0
	Phasianus	*colchicus*	306	0
Gaviiformes	*Gavia*	*arctica*	1	0
	Gavia	*pacifica*	1	0
	Gavia	*stellata*	3	0
Gruiformes	*Amaurornis*	*phoenicurus*	17	1
	Fulica	*americana*	59	1
	Fulica	*atra*	428	12
	Gallicrex	*cinerea*	117	15
	Gallinula	*chloropus*	168	8
	Gallirallus	*striatus*	49	8
	Otis	*tarda*	14	0
	Porphyrio	*porphyrio*	12	0
	Porzana	*carolina*	1	0
	Porzana	*cinerea*	8	1

Appendix B continued

Order	Genus	Species	# Samples	# AIV Pos
Gruiformes	*Porzana*	*fusca*	498	53
	Porzana	*parva*	2	0
	Porzana	*porzana*	11	0
	Porzana	*pusilla*	9	1
	Rallus	*aquaticus*	3	0
	Rallus	*elegans*	1	0
Passeriformes	*Acridotheres*	*grandis*	49	5
	Acridotheres	*tristis*	77	14
	Acrocephalus	*aedon*	4	0
	Acrocephalus	*agricola*	1	0
	Acrocephalus	*bistrigiceps*	8	0
	Acrocephalus	*orientalis*	62	3
	Acrocephalus	*spp.*	2	0
	Aegithina	*tiphia*	8	1
	Anthus	*hodgsoni*	1	0
	Anthus	*lutescens*	1	0
	Anthus	*rubescens*	2	0
	Anthus	*spinoletta*	1	0
	Calcarius	*lapponicus*	1	0
	Carduelis	*flammea*	17	0
	Catharus	*guttatus*	1	0
	Catharus	*ustulatus*	27	1
	Copsychus	*saularis*	6	0
	Corvus	*brachyrhynchos*	36	0
	Corvus	*corax*	2	0
	Corvus	*corone cornix*	3	0
	Corvus	*frugilegus*	7	0
	Crypsirina	*temia*	8	3
	Cyornis	*hainanus*	1	0
	Dendroica	*coronata*	17	0
	Dendroica	*petechia*	10	3
	Dicrurus	*hottentottus*	1	0
	Dicrurus	*macrocercus*	4	0
	Dives	*dives*	2	0
	Dumetella	*carolinensis*	45	0
	Emberiza	*aureola*	20	0
	Emberiza	*chrysophrys*	1	0
	Emberiza	*fucata*	1	0
	Emberiza	*pallasi*	8	0
	Emberiza	*rutila*	5	0
	Emberiza	*spodocephala*	20	0
	Empidonax	*hammondii*	7	0
	Euphagus	*carolinus*	2	0

Appendix B continued

Order	Genus	Species	# Samples	# AIV Pos
Passeriformes	*Euphonia*	*hirundinacea*	1	0
	Ficedula	*parva*	11	1
	Garrulus	*glandarius*	1	0
	Gracula	*religiosa*	3	0
	Hirundo	*rustica*	79	3
	Hylocichla	*mustelina*	1	0
	Hypothymis	*azurea*	7	1
	Icteria	*virens*	4	0
	Icterus	*galbula*	8	0
	Icterus	*spurius*	1	0
	Junco	*hyemalis*	11	0
	Lanius	*collurio*	4	0
	Lanius	*cristatus*	16	1
	Lanius	*schach*	1	0
	Locustella	*certhiola*	23	0
	Lonchura	*malacca*	32	2
	Lonchura	*punctulata*	143	8
	Lonchura	*striata*	4	0
	Luscinia	*calliope*	15	2
	Luscinia	*cyane*	1	0
	Luscinia	*svecica*	2	0
	Macronous	*gularis*	3	0
	Megalurus	*palustris*	1	0
	Melospiza	*lincolnii*	3	0
	Mionectes	*oleaginus assimilis*	1	0
	Motacilla	*alba*	4	0
	Motacilla	*alba ocularis*	1	0
	Motacilla	*taivana*	2	0
	Motacilla	*tschutschensis*	13	0
	Muscicapa	*dauurica*	5	0
	Myiarchus	*crinitus*	1	0
	Myiozetetes	*similis*	1	0
	Orthotomus	*atrogularis*	1	0
	Oryzoborus	*funereus*	9	0
	Padda	*oryzivora*	4	0
	Passer	*domesticus*	28	2
	Passer	*flaveolus*	52	1
	Passer	*montanus*	141	27
	Passerculus	*sandwichensis*	2	0
	Passerina	*cyanea*	4	0
	Pellorneum	*ruficeps*	1	0
	Phylloscopus	*borealis*	4	0
	Phylloscopus	*fuscatus*	2	0

Appendix B continued

Order	Genus	Species	# Samples	# AIV Pos
Passeriformes	*Phylloscopus*	*schwarzi*	3	1
	Phylloscopus	*trochiloides*	2	0
	Piranga	*rubra*	1	0
	Pitangus	*sulphuratus*	2	1
	Ploceus	*hypoxanthus*	11	1
	Ploceus	*manyar*	46	4
	Ploceus	*philippinus*	125	3
	Poecile	*atricapillus*	1	0
	Prinia	*flaviventris*	1	0
	Prinia	*inornata*	9	5
	Prinia	*polychroa*	1	0
	Psarocolius	*montezuma*	1	0
	Pycnonotus	*aurigaster*	8	2
	Pycnonotus	*blanfordi*	85	16
	Pycnonotus	*finlaysoni*	16	3
	Pycnonotus	*goiavier*	14	3
	Quiscalus	*mexicanus*	26	0
	Rhipidura	*javanica*	23	9
	Riparia	*riparia*	1	0
	Saxicola	*caprata*	1	0
	Saxicola	*maura*	1	0
	Saxicola	*torquata*	54	3
	Scaphidura	*oryzivora*	5	0
	Seiurus	*aurocapillus*	3	0
	Seiurus	*noveboracensis*	56	10
	Sporophila	*aurita*	3	0
	Sporophila	*torqueola*	27	5
	Sturnus	*nigricollis*	7	2
	Sturnus	*vulgaris*	14	0
	Tachycineta	*bicolor*	1	0
	Thraupis	*episcopus*	2	0
	Thryothorus	*maculipectus*	2	0
	Timalia	*pileata*	3	1
	Turdus	*eunomus*	10	0
	Turdus	*grayi*	48	0
	Turdus	*merula*	6	0
	Turdus	*migratorius*	24	0
	Turdus	*naumanni*	1	0
	Turdus	*pallidus*	2	0
	Turdus	*viscivorus*	4	0
	Vermivora	*celata*	2	0
	Vireo	*griseus*	2	0
	Vireo	*olivaceus*	1	0

Appendix B continued

Order	Genus	Species	# Samples	# AIV Pos
Passeriformes	*Volatinia*	*jacarina*	3	0
	Wilsonia	*pusilla*	1	0
	Zonotrichia	*leucophrys*	2	0
Pelecaniformes	*Ardea*	*alba*	3	0
	Ardea	*cinerea*	73	4
	Ardea	*herodias*	3	0
	Ardea	*purpurea*	76	0
	Ardeola	*bacchus*	2	1
	Ardeola	*bacchus/speciosa*	925	114
	Butorides	*striatus*	2	0
	Butorides	*virescens*	14	0
	Casmerodius	*albus*	1	0
	Egretta	*alba*	2	0
	Egretta	*garzetta*	55	3
	Egretta	*gularis*	17	0
	Egretta	*thula*	142	0
	Ixobrychus	*cinnamomeus*	22	5
	Ixobrychus	*sinensis*	34	1
	Ixobrychus	*spp.*	4	0
	Mesophoyx	*intermedia*	25	1
	Nycticorax	*nycticorax*	228	3
	Pelecanus	*erythrorhynchos*	39	0
	Pelecanus	*occidentalis*	9	0
	Pelecanus	*philippensis*	1	0
	Threskiornis	*melanocephalus*	2	1
Piciformes	*Celeus*	*brachyurus*	1	0
	Dinopium	*javanense*	3	0
	Melanerpes	*aurifrons*	15	1
	Sphyrapicus	*varius*	1	0
Podicipediformes	*Aechmophorus*	*occidentalis*	10	0
	Podiceps	*auritus*	1	0
	Podiceps	*cristatus*	5	0
	Podiceps	*nigricollis*	8	0
	Podilymbus	*podiceps*	4	0
	Tachybaptus	*ruficollis*	9	0
Procellariiformes	*Fulmarus*	*glacialis*	2	0
	Oceanodroma	*homochroa*	1	0
	Puffinus	*griseus*	55	0
Psittaciformes	*Ara*	*macao*	24	0
	Cacatua	*galerita*	24	0
	Melopsittacus	*undulatus*	9	0
	Psittacula	*krameri*	3	0
Strigiformes	*Asio*	*flammeus*	1	0

Appendix B continued

Order	Genus	Species	# Samples	# AIV Pos
Strigiformes	*Asio*	*otus*	6	0
	Bubo	*virginianus*	69	0
	Glaucidium	*brasilianum*	2	1
	Glaucidium	*brodiei*	1	0
	Otus	*bakkamoena*	1	0
	Tyto	*alba*	12	0
Suliformes	*Anhinga*	*melanogaster*	1	0
	Phalacrocorax	*auritus*	7	0
	Phalacrocorax	*carbo*	210	31
	Phalacrocorax	*niger*	5	1
	Phalacrocorax	*pelagicus*	1	0
	Phalacrocorax	*penicillatus*	15	0
	Phalacrocorax	*pygmeus*	19	0

Appendix C. List of bird species in the Alaska Asia Avian Influenza Research 2005-2010 database. The database was analyzed in Chapter 4 and this list includes order, scientific name, total number of samples collected for each species (# Samples), and number of samples testing positive for avian influenza virus (# AIV Pos) for each species. This dataset contains 75 species of bird.

Order	***Genus***	***Species***	**# Samples**	**# AIV Pos**
Anseriformes	*Anas*	*acuta*	15277	724
	Anas	*americana*	3202	31
	Anas	*clypeata*	2662	83
	Anas	*crecca*	3031	108
	Anas	*discors*	33	0
	Anas	*platyrhynchos*	9321	556
	Anas	*platyrhynchos*	1	0
	Anas	*strepera*	176	3
	Anas	*strepera*	2	0
	Anser	*albifrons*	9	0
	Aythya	*affinis*	2606	15
	Aythya	*americana*	10	0
	Aythya	*collaris*	130	2
	Aythya	*marila*	252	4
	Aythya	*spp.*	72	0
	Aythya	*valisineria*	148	1
	Branta	*canadensis*	2	0
	Branta	*canadensis*	280	0
	Bucephala	*albeola*	72	0
	Bucephala	*clangula*	22	0
	Bucephala	*islandica*	17	2
	Clangula	*hyemalis*	7	0
	Melanitta	*fusca*	1	0
	Somateria	*spectabilis*	11	0
Charadriiformes	*Calidris*	*mauri*	321	7
	Calidris	*minutilla*	94	0
	Calidris	*pusilla*	3	0
	Calidris	*sp.*	5	0
	Fratercula	*cirrhata*	1	0
	Larus	*(sp)*	16	0
	Larus	*canus*	21	0
	Larus	*glaucescens*	549	9
	Larus	*philadelphia*	1	0
	Phalaropus	*fulicarius*	3	0
	Phalaropus	*lobatus*	2	0
	Tringa	*solitaria*	2	0
	Uria	*aalge*	3	0
Galliformes	*Gallus*	*gallus*	4	0
Gaviiformes	*Gavia*	*adamsii*	2	0

Appendix C continued

Order	*Genus*	*Species*	# Samples	# AIV Pos
Passeriformes	*Carduelis*	*flammea*	189	0
	Catharus	*guttatus*	18	0
	Catharus	*minimus*	9	0
	Catharus	*ustulatus*	84	1
	Dendroica	*coronata*	446	5
	Dendroica	*petechia*	72	1
	Dendroica	*striata*	5	0
	Dendroica	*townsendi*	4	0
	Empidonax	*alnorum*	15	0
	Empidonax	*hammondii*	101	5
	Euphagus	*carolinus*	12	0
	Hylocichla	*mustelina*	5	0
	Ixoreus	*naevius*	4	1
	Junco	*hyemalis*	1	0
	Junco	*hyemalis*	276	4
	Melospiza	*lincolnii*	189	7
	Oreothlypis	*celata*	195	9
	Parkesia	*noveboracensis*	56	0
	Passerculus	*sandwichensis*	27	0
	Passerella	*iliaca*	18	0
	Phylloscopus	*borealis*	1	0
	Poecile	*atricapillus*	119	3
	Poecile	*hudsonica*	13	0
	Regulus	*calendula*	21	0
	Regulus	*satrapa*	4	0
	Riparia	*riparia*	1	0
	Spizella	*arborea*	36	0
	Tachycineta	*bicolor*	185	1
	Tachycineta	*thalassina*	4	0
	Turdus	*migratorius*	87	1
	Wilsonia	*pusilla*	35	0
	Zonotrichia	*leucophrys*	21	0
Piciformes	*Picoides*	*pubescens*	2	0
Procellariiformes	*Oceanodroma*	*furcata*	31	0
	Oceanodroma	*leucorhoa*	70	0
	Oceanodroma	*tristrami*	1	0

Appendix D. List of bird species in the Canada's Inter-agency Wild Bird Influenza survey (CIWBI) database. The database was analyzed in Chapter 4 and this list includes order, scientific name, total number of samples collected for each species (# Samples), and number of samples testing positive for avian influenza virus (# AIV Pos) for each species. This dataset contains 189 species of bird.

Order	Genus	Species	# Samples	# AIV Pos
Accipitriformes	*Accipiter*	*cooperii*	13	0
	Accipiter	*gentilis*	5	0
	Accipiter	*striatus*	22	0
	Accipitridae	*spp.*	1	0
	Aquila	*chrysaetos*	6	0
	Cathartes	*aura*	1	0
Anseriformes	*Aix*	*sponsa*	111	36
	Anas	*acuta*	211	74
	Anas	*americana*	11	0
	Anas	*carolinensis*	11	2
	Anas	*clypeata*	6	1
	Anas	*discors*	1405	399
	Anas	*platyrhynchos*	533	89
	Anas	*platyrhynchos domesticus*	2	0
	Anas	*rubripes*	2	1
	Anas	*spp.*	14	1
	Anas	*strepera*	4	0
	Anser	*albifrons*	1	0
	Aythya	*affinis*	9	2
	Aythya	*americana*	80	35
	Aythya	*marila*	2	0
	Aythya	*valisineria*	19	15
	Branta	*canadensis*	69	0
	Bucephala	*albeola*	1	0
	Bucephala	*clangula*	2	0
	Bucephala	*islandica*	1	0
	Chen	*caerulescens*	176	0
	Chen	*rossii*	1	0
	Cygnus	*buccinator*	99	2
	Cygnus	*columbianus*	9	0
	Cygnus	*olor*	5	0
	Cygnus	*spp.*	1	0
	Histrionicus	*histrionicus*	1	0
	Lophodytes	*cucullatus*	3	0
	Melanitta	*perspicillata*	2	0
	Oxyura	*jamaicensis*	2	0
Apodiformes	*Archilochus*	*colubris*	3	0
	Selasphorus	*rufus*	1	0
Caprimulgiformes	*Chordeiles*	*minor*	1	0

Appendix D continued

Order	Genus	Species	# Samples	# AIV Pos
Charadriiformes	*Brachyramphus*	*marmoratus*	1	0
	Calidris	*alpina*	1	0
	Calidris	*bairdii*	1	0
	Calidris	*pusilla*	1	0
	Cepphus	*columba*	7	0
	Cerorhinca	*monocerata*	11	0
	Charadrius	*melodus*	2	0
	Laridae	*spp.*	14	0
	Larus	*argentatus*	19	0
	Larus	*californicus*	1	0
	Larus	*delawarensis*	32	5
	Larus	*fuscus*	1	0
	Larus	*glaucescens*	4	0
	Larus	*marinus*	3	0
	Larus	*pipixcan*	17	0
	Phalaropus	*lobatus*	1	0
	Ptychoramphus	*aleuticus*	1	0
	Scolopacidae	*spp.*	1	0
	Sterna	*caspia*	2	0
	Tryngites	*subruficollis*	1	0
	Uria	*aalge*	25	0
	Uria	*lomvia*	52	0
Ciconiiformes	*Botaurus*	*lentiginosus*	1	0
Columbiformes	*Columba*	*livia*	76	0
	Zenaida	*macroura*	12	0
Coraciiformes	*Ceryle*	*alcyon*	2	0
Falconiformes	*Buteo*	*jamaicensis*	38	0
	Buteo	*lagopus*	6	0
	Buteo	*platypterus*	2	0
	Buteo	*regalis*	2	0
	Buteo	*swainsoni*	18	0
	Circus	*cyaneus*	5	0
	Falco	*biarmicus*	1	0
	Falco	*columbarius*	29	0
	Falco	*peregrinus*	11	0
	Falco	*rusticolus*	2	0
	Falco	*sparverius*	4	0
	Haliaeetus	*leucocephalus*	82	0
	Pandion	*haliaetus*	1	0
Galliformes	*Bonasa*	*umbellus*	3	0
	Callipepla	*californica*	2	0

Appendix D continued

Order	Genus	Species	# Samples	# AIV Pos
Galliformes	*Meleagris*	*gallopavo*	6	0
Gaviiformes	*Gavia*	*immer*	16	0
	Gavia	*pacifica*	2	0
	Gavia	*stellata*	1	0
Gruiformes	*Fulica*	*americana*	19	1
	Grus	*canadensis*	1	0
	Porzana	*carolina*	1	0
Passeriformes	*Agelaius*	*phoeniceus*	2	0
	Bombycilla	*cedrorum*	13	0
	Bombycilla	*garrulus*	21	0
	Carduelis	*flammea*	1	0
	Carduelis	*pinus*	2	0
	Carduelis	*tristis*	3	0
	Carpodacus	*mexicanus*	6	0
	Carpodacus	*purpureus*	1	0
	Catharus	*guttatus*	2	0
	Catharus	*ustulatus*	4	0
	Certhia	*americana*	1	0
	Contopus	*borealis*	1	0
	Corvus	*brachyrhynchos*	138	0
	Corvus	*caurinus*	7	0
	Corvus	*corax*	19	0
	Cyanocitta	*cristata*	15	0
	Cyanocitta	*stelleri*	1	0
	Dendroica	*coronata*	3	0
	Dendroica	*fusca*	2	0
	Dendroica	*magnolia*	1	0
	Dendroica	*pensylvanica*	2	0
	Dendroica	*petechia*	4	0
	Dendroica	*striata*	11	0
	Dendroica	*virens*	1	0
	Dumetella	*carolinensis*	1	0
	Emberizidae	*spp.*	1	0
	Empidonax	*minimus*	1	0
	Euphagus	*cyanocephalus*	3	0
	Geothlypis	*trichas*	1	0
	Hirundo	*rustica*	3	0
	Hylocichla	*mustelina*	1	0
	Icteridae	*spp.*	10	0
	Icterus	*galbula*	3	0
	Icterus	*spurius*	1	0
	Junco	*hyemalis*	6	0
	Lanius	*excubitor*	1	0

Appendix D continued

Order	Genus	Species	# Samples	# AIV Pos
Passeriformes	*Lanius*	*ludovicianus*	54	0
	Loxia	*curvirostra*	1	0
	Loxia	*leucoptera*	3	0
	Melospiza	*lincolnii*	2	0
	Melospiza	*melodia*	3	0
	Molothrus	*ater*	3	0
	Myiarchus	*crinitus*	1	0
	Parulidae	*spp.*	5	0
	Parus	*atricapillus*	4	0
	Passer	*domesticus*	30	0
	Passeriformes	*spp.*	5	0
	Phainopepla	*nitens*	1	0
	Pheucticus	*ludovicianus*	9	0
	Pica	*hudsonia*	12	0
	Plectrophenax	*nivalis*	1	0
	Progne	*subis*	7	0
	Quiscalus	*quiscula*	25	0
	Regulus	*satrapa*	4	0
	Seiurus	*aurocapillus*	3	0
	Seiurus	*noveboracensis*	1	0
	Setophaga	*ruticilla*	2	0
	Sialia	*sialis*	1	0
	Sitta	*canadensis*	1	0
	Spizella	*arborea*	1	0
	Spizella	*passerina*	6	0
	Sturnella	*neglecta*	1	0
	Sturnidae	*spp.*	1	0
	Sturnus	*vulgaris*	7	0
	Tachycineta	*bicolor*	15	0
	Toxostoma	*rufum*	1	0
	Troglodytes	*aedon*	2	0
	Turdus	*merula*	1	0
	Turdus	*migratorius*	85	0
	Tyrannus	*tyrannus*	1	0
	Vermivora	*celata*	2	0
	Vermivora	*peregrina*	7	0
	Vermivora	*ruficapilla*	1	0
	Vireo	*olivaceus*	2	0
	Wilsonia	*canadensis*	10	0
	Wilsonia	*pusilla*	1	0
	Zonotrichia	*leucophrys*	1	0
Pelecaniformes	*Ardea*	*herodias*	15	0
	Nycticorax	*nycticorax*	2	0

Appendix D continued

Order	Genus	Species	# Samples	# AIV Pos
Pelecaniformes	*Pelecanus*	*erythrorhynchos*	16	0
Piciformes	*Colaptes*	*auratus*	27	0
	Picidae	*spp.*	1	0
	Picoides	*villosus*	1	0
	Sphyrapicus	*varius*	4	0
Podicipediformes	*Aechmophorus*	*occidentalis*	1	0
	Podiceps	*grisegena*	2	0
Podicipediformes	*Podiceps*	*nigricollis*	1	0
	Podilymbus	*podiceps*	2	0
Strigiformes	*Aegolius*	*acadicus*	6	0
	Aegolius	*funereus*	1	0
	Asio	*flammeus*	5	0
	Asio	*otus*	4	0
	Athene	*cunicularia*	8	0
	Bubo	*virginianus*	62	0
	Nyctea	*scandiaca*	2	0
	Strix	*nebulosa*	2	0
	Strix	*occidentalis*	2	0
	Strix	*varia*	30	0
	Surnia	*ulula*	1	0
	Tyto	*alba*	17	0
Suliformes	*Phalacrocorax*	*auritus*	69	0
	Phalacrocorax	*pelagicus*	4	0